Semi-automated co-reference identification in digital humanities collections.

David Croft

Submitted in partial fulfilment of
the requirements for the degree of Doctor of Philosophy
at De Montfort University

July 16, 2014

Acknowledgements

This thesis would not have been possible without the help and support of my family, friends and colleagues. I would like to give special thanks to:

Jo, whose support during this project has been boundless and whose nagging about finishing this thesis wasn't annoying at all.

My Mum, Heather, and Dad, Jeremy. Who have loved, supported and encouraged me over the years and will now hopefully forget about the project which shot across the living room floor and caught fire.

My supervisors, Stephen Brown and Simon Coupland. Without whose support, guidance and enthusiasm for my research, this thesis would not have been possible and would have contained a lot of spelling mistaeks.

The staff and students of the Photographic History Research Centre at De Montfort University. Particularly Kelly Wilder, for her invaluable insights into photohistory and the GLAM community.

Roger Taylor for his excellent explanation of the relationships between photographic processes.

Finally I would like to thank De Montfort University for its support, and the Arts and Humanities Research Council (AHRC) for funding this research.

Contents

Contents **i**

1 Introduction **1**
- 1.1 Research focus . . . 3
- 1.2 The GLAM community's expectations . . . 8
  - 1.2.1 Field use/importance . . . 9
  - 1.2.2 Number of results expected/desired . . . 11
  - 1.2.3 Perceived effectiveness of the current search systems . . . 11
  - 1.2.4 Search strategy . . . 12
  - 1.2.5 Conclusion . . . 14
- 1.3 Thesis layout . . . 14

2 Query expansion **17**
- 2.1 Spelling correction . . . 19
- 2.2 Stemming/term expansion . . . 21
- 2.3 Synonym expansion . . . 22
  - 2.3.1 Global reference approach . . . 23
  - 2.3.2 Relevance feedback approach . . . 29
  - 2.3.3 Comparison of synonym expansion approaches . . . 30
- 2.4 Conclusions . . . 33

3 Co-reference identification **34**
- 3.1 Rule based identification . . . 37
  - 3.1.1 Expert systems . . . 38
  - 3.1.2 Fuzzy Logic . . . 40
- 3.2 Probabilistic Record Linkage (PRL) . . . 45
- 3.3 Artificial Neural Networks (ANNs) . . . 48
- 3.4 Clustering . . . 50

3.4.1 Hierarchical clustering . . . 50
3.4.2 Partitional . . . 51
3.4.3 Description of k-means . . . 52
3.4.4 k requirement . . . 54
3.4.5 Density based clustering . . . 56
3.4.6 Hard vs. fuzzy . . . 56
3.4.7 Post clustering processing . . . 57
3.5 Conclusions . . . 57

4 Text comparison **60**
4.1 Approximate string comparison . . . 60
4.1.1 Phonetic . . . 62
4.1.2 Edit distance . . . 64
4.1.3 Edit distance resembling approaches . . . 66
4.1.4 Conclusions . . . 68
4.2 Textual similarity . . . 69
4.2.1 Term Frequency (TF) and TF-Inverse Document Frequency (IDF) . . . 70
4.2.2 Binary vector methods . . . 71
4.2.3 Cosine similarity . . . 72
4.2.4 Okapi BM25F . . . 73
4.2.5 Conclusions . . . 74
4.3 Semantic string comparison . . . 74
4.3.1 Latent Semantic Analysis (LSA) . . . 75
4.3.2 STASIS . . . 76
4.4 Conclusions . . . 78

5 Collections **80**
5.1 Markup languages and metadata schemas . . . 81
5.2 Ontologies . . . 83
5.3 Syntax independence . . . 84
5.4 Resource access . . . 85
5.5 Conclusion . . . 86

6 Methodology **87**
6.0.1 Lack of image processing . . . 90
6.1 Keyword extraction/expansion . . . 91

6.2 Searching external collections . . . 93
6.2.1 Simulating collection Application Programming Interfaces (APIs) . . . 94
6.3 Individual field processing . . . 98
6.3.1 De-duplication of field values . . . 98
6.3.2 *Title* field metric . . . 100
6.3.3 *Description* field metric . . . 105
6.3.4 *Person* field metric . . . 106
6.3.5 *Process* field metric . . . 110
6.3.6 *Date* field metric . . . 115
6.4 Overall record similarity . . . 119
6.4.1 Fuzzy Inference System (FIS) . . . 120
6.5 Record ordering . . . 122
6.5.1 Clustering . . . 123
6.5.2 Constrained search . . . 124
6.6 Summary . . . 126

7 Testing **128**
7.1 Result quality . . . 130
7.1.1 Data collection . . . 132
7.1.2 Analysis . . . 138
7.1.3 Testing problems . . . 146
7.1.4 Conclusions . . . 151
7.2 *Title* metric testing . . . 151
7.2.1 Data collection . . . 152
7.2.2 Analysis . . . 154
7.3 Collections searched . . . 155
7.3.1 Data collection . . . 155
7.3.2 Analysis/conclusions . . . 156
7.4 Time taken . . . 157
7.5 Conclusions . . . 158

8 Conclusions **161**
8.0.1 Main contributions . . . 165
8.1 With the benefit of hindsight . . . 165
8.2 Further work . . . 166

Bibliography **170**

A Collection records **190**
A.1 Brooklyn Museum (BkM) . . . 190
A.1.1 Example data . . . 190
A.2 DigitalNZ (New Zealand) (DNZ) . . . 192
A.2.1 Example data . . . 192
A.3 Exhibitions of the Royal Photographic Society (ERPS) . . . 194
A.4 Library of Congress (LoC) . . . 194
A.4.1 Example data . . . 194
A.5 Photographic Exhibitions in Britain (PEiB) . . . 196
A.6 Victoria and Albert Museum (V&A) . . . 196
A.6.1 Example data . . . 197

B *Title* field **206**

C *Person* field **208**

D *Process* field **210**
D.1 Types . . . 210
D.2 Groups . . . 211

E Overall record **213**

F Testing participant responses **215**

G Potential co-reference matches **217**

H Potential matches for the 'missing' ERPS photographs **221**

I Questionnaires **227**
I.1 Search technique questionnaire . . . 227
I.2 User testing questionnaire . . . 231

J Journal/conference papers **237**

List of Figures

1.1 The ERPS collection website showing record erps17654. 4
1.2 Example section from the original ERPS catalogues. 5
1.3 Example of a sketched photograph, erps17094. 5
1.4 Comparison of field use and importance as reported by twenty three respondents from the Gallery, Library, Archive and Museum (GLAM) community. 10
1.5 The number of results for a search query that are expected, desired and tolerated As reported by a sample of the GLAM community. . . 11
1.6 Precision and recall expectations of current search systems. As reported by a sample of the GLAM community. 12
1.7 The relative use of different search techniques by a sample of the GLAM community. 13
1.8 Number of collections regularly accessed by questionnaire participants. 13

2.1 Visual demonstration of precision and recall in searching. 18

3.1 Input and output sets for and example FIS. 42
3.2 Visualisation of an example FIS. 42
3.3 Defuzzification methods for fuzzy sets. 43
3.4 Centroid defuzzification. 44
3.5 Weighted average defuzzification. 44
3.6 Example of interconnected nodes in a MultiLayer Perceptron (MLP) ANN. 49
3.7 Example of a dendrogram which could be produced by hierarchical clustering. 51
3.8 Example of poor initial centroid placement for k-means clustering. . . 52
3.9 Example k-means clustering on concave clusters. 53
3.10 Example of an unsuitable k value for k-means clustering. 54

3.11 Example of k-means clustering on clusters of difference sizes/densities. 54
3.12 Example data points and corresponding Visual Assessment of cluster Tendency (VAT) image. 55

4.1 Plotted term vectors from table 4.2. 72
4.2 Initial matrices used in LSA. 75
4.3 Truncation of Singular Value Decomposition (SVD) matrices. 76
4.4 Truncated/optimised matrices used in LSA. 76

5.1 Example of the Dublin Core schema marked up in eXtensible Markup Language (XML). 81
5.2 Typical example of the Dublin Core schema marked up in XML, following XML conventions and recommendations. 82
5.3 Example of the Dublin Core schema marked up in JavaScript Object Notation (JSON). 82

6.1 Process flow diagram for the proposed approach. 88
6.2 The actions needed in order to search through three collections using traditional search interfaces. 89
6.3 Manually performed actions to search through three collections using the proposed approach. 90
6.4 Comparison of record counts from ERPS seeded searches versus the number of unique values per field. 99
6.5 Example of the difference between term vectors and weighted vectors. 102
6.6 Subset of the processes hierarchy used by the *process* metric. 113
6.7 Example date ranges and gaps. 119
6.8 Fuzzy sets used for overall record similarity. 121
6.9 FIS output surface, inputs *title* and *person*, *process* = *date* = 0.0. . . 122
6.10 FIS output surface, inputs *title* and *process*/*date*, *person* = 0.0. . . . 122
6.11 FIS output surface, inputs *title*/*person* and *process*/*date*. 123
6.12 VAT images of various erps17093 similarity matrices. 124
6.13 Example result, full graph for ERPS record 17093. 125
6.14 Example result, top 100 results for ERPS record 17093. 126
6.15 Example result, top 4 results for ERPS record 17093. 126

7.1 Top records of the erps28409 dendrogram. 131

7.2 Percentage of tests in which each approach was deemed to have found a co-reference match. 139
7.3 Percentage of distinct records for which each approach was deemed to have found a co-reference match. 139
7.4 Match quality ratings for erps17093 given by the testing participants. 142
7.5 Match quality ratings for erps28409 given by the testing participants. 143
7.6 Approach preferred by test participants on a per search basis. 145
7.7 Responses to the question "If the test approach was made available to you, would you use it for searching in the future?" as percentages. 146
7.8 Combined field similarity matrix sizes verses predicted size. 148
7.9 Memory requirements for proposed approach. 149
7.10 Distributions for the number of records returned with and without keyword filtering. 150
7.11 Human, LSA, STASIS and *title* metric generated similarity values for STSS-65 subset. 153
7.12 Comparison of processing time requirements for LSA vs. *title* metric vs. STASIS. 155
7.13 Collections used. 156

D.1 Network diagram showing the individual photographic processes, their keywords and relationships as understood by the *process* metric. 212

List of Tables

2.1 Example use of cosine similarity on a corpus for identifying synonyms. 26

3.1 Example of training and co-reference identification using PRL. 47

4.1 Character values for the Soundex algorithm. 63
4.2 Term vectors for cosine example. 72

6.1 Estimates of overall collection size and number of records actually collected for this research. 98
6.2 Example term similarity matrix 104
6.3 Example of original and corresponding weighted term vectors. 105
6.4 Jaro-Winkler similarity matrix. 108
6.5 Ordered Jaro-Winkler similarity matrix. 108
6.6 *Person* field similarity metric result. 109
6.7 Example comparison of *process* field containing 'platinotypes' against a subsection of process keyword sets. 114
6.8 Process types similarity values. 114

7.1 Co-reference candidates for erps28409. 132
7.2 Co-reference candidates for erps17093. 137
7.3 Example of a promising but non-matching co-reference candidate for erps28409. 141
7.4 Average rankings for the pre-selected test records. 144
7.5 Mann-Whitney U (MWU) test results. 144
7.6 Average number of unique values per record. 147
7.7 Time spent searching by test participants. 158

B.1 Raw results for *title* metric testing using STSS-65. 207

F.1 Occurrences of co-referent matches per search approach. 215

F.2 Participant search approach result rankings. 216

G.1 Co-reference candidates for erps16545. 217
G.2 Co-reference candidates for erps16578. 218
G.3 Co-reference candidates for erps16939. 219
G.4 Co-reference candidates for erps18559. 219
G.5 Co-reference candidates for erps18912. 220

H.1 Co-reference candidates for an ERPS record with no image (erps8122). 221
H.2 Co-reference candidates for an ERPS record with no image (erps15874). 222
H.3 Co-reference candidates for an ERPS record with no image (erps18923). 222
H.4 Co-reference candidates for an ERPS record with no image (erps19389). 223
H.5 Co-reference candidates for an ERPS record with no image (erps19533). 223
H.6 Co-reference candidates for an ERPS record with no image (erps25184). 224
H.7 Co-reference candidates for an ERPS record with no image (erps26130). 224
H.8 Co-reference candidates for an ERPS record with no image (erps32607). 225
H.9 Co-reference candidates for an ERPS record with no image (erps32622). 225
H.10 Co-reference candidates for an ERPS record with no image (erps33446). 226

List of Algorithms

1 Levenshtein distance. 67
2 Algorithm for the *title* similarity metric. 206
3 Algorithm for the *person* similarity metric. 208
4 Constrained minimum spanning tree algorithm. 214

Acronyms

AI Artificial Intelligence. 20, 37, 138
ANN Artificial Neural Network. 3, 13, 35, 44–46, 101
API Application Programming Interface. 75, 79, 80, 118, 137, 141, 142

BkM Brooklyn Museum. 80, 82, 113, 118, 141, 153

CI Computational Intelligence. 3, 5, 137
CLIQUE CLustering In QUEst. 52
COPSY Context OPerator SYntax. 24
CSV Comma Separated Values. 57

DBSCAN Density Based Spatial Clustering of Applications with Noise. 52
DENCLUE DENsity CLUstEring. 52
DENDRAL DENDRitic ALgorithm. 37
DMU De Montfort University. 3, 4
DNZ DigitalNZ (New Zealand). 80, 82, 113, 141

ERPS Exhibitions of the Royal Photographic Society. 3–5, 8, 9, 18, 20, 32, 34, 53, 54, 57, 64, 65, 69, 70, 74, 75, 77, 80–83, 89, 91, 106–108, 113, 114, 118, 123, 124, 126, 132–135, 137, 138, 141, 142

FIS Fuzzy Inference System. 35, 38–42, 76, 101–104
FLAME Fuzzy clustering by Local Approximation of MEmberships. 52

GLAM Gallery, Library, Archive and Museum. 1–15, 17, 18, 20, 26–32, 34, 35, 43, 45, 46, 53, 54, 56, 60, 64, 69, 74, 75, 78–80, 84, 89, 93, 99–103, 107, 110, 123, 127, 136–142
GUI Graphical User Interface. 36
GUID Globally Unique IDentifier. 32, 100

HTML HyperText Markup Language. 57

IDF Inverse Document Frequency. 65, 66
IMLS Institute of Museum and Library Services. 1, 2
ISBN International Standard Book Number. 32, 33

JSON JavaScript Object Notation. 69

LDB Lexical DataBase. 21, 22, 25–28, 71, 77
LoC Library of Congress. 80, 82, 90, 91, 98, 113, 114, 133, 141
LOD Linking Open Data. 7
LSA Latent Semantic Analysis. 70–72, 74, 83, 84, 128–132, 135, 139, 145
LSI Latent Semantic Indexing. 70
LSQ Least SQuares. 42

MAFIA Merging of Adaptive Finite IntervAls. 52
MeSH Medical Subject Headings. 21, 26
MLP MultiLayer Perceptron. 44, 46
MST Minimum Spanning Tree. 105
MWU Mann-Whitney U. 121, 122

NI National Insurance. 32
NLP Natural Language Processing. 24, 29, 83

PEIB Photographic Exhibitions in Britain. 81, 82, 113, 133, 141
PRL Probabilistic Record Linkage. 13, 42–45, 54, 101

RBF Radial Basis Function. 44
RDF Resource Description Framework. 7
REST REpresentational State Transfer. 79, 80, 118, 153
RNN Recurrent Neural Network. 44
RPS Royal Photographic Society. 3

SEXTANT Semantic EXtraction from Text via Analyzed Networks of Terms. 24, 25
SPARQL SPARQL Protocol and Rdf Query Language. 7, 79, 80
SPSS IBM Statistical Product and Service Solutions. 121
SSCI Symposium Series on Computational Intelligence. 166
STSS Short Text Semantic Similarity. 70, 135, 138, 139
SVD Singular Value Decomposition. 70, 71

TF Term Frequency. 65–67, 83

TF-IDF Term Frequency-Inverse Document Frequency. 66, 70, 87, 88
TSK Takagi-Sugeno-Kang. 38, 41

UID Unique IDentifier. 32, 33

V&A Victoria and Albert Museum. 79–82, 113, 114, 133, 141
VAT Visual Assessment of cluster Tendency. 51, 52, 101, 104, 105

Web World Wide Web. 1, 25, 26
WSD Word Sense Disambiguation. 29

XML eXtensible Markup Language. 69

Abstract

Locating specific information within museum collections represents a significant challenge for collection users. Even when the collections and catalogues exist in a searchable digital format, formatting differences and the imprecise nature of the information to be searched mean that information can be recorded in a large number of different ways. This variation exists not just between different collections, but also within individual ones. This means that traditional information retrieval techniques are badly suited to the challenges of locating particular information in digital humanities collections and searching, therefore, takes an excessive amount of time and resources.

This thesis focuses on a particular search problem, that of co-reference identification. This is the process of identifying when the same real world item is recorded in multiple digital locations. In this thesis, a real world example of a co-reference identification problem for digital humanities collections is identified and explored. In particular the time consuming nature of identifying co-referent records. In order to address the identified problem, this thesis presents a novel method for co-reference identification between digitised records in humanities collections. Whilst the specific focus of this thesis is co-reference identification, elements of the method described also have applications for general information retrieval.

The new co-reference method uses elements from a broad range of areas including; query expansion, co-reference identification, short text semantic similarity and fuzzy logic. The new method was tested against real world collections information, the results of which suggest that, in terms of the quality of the co-referent matches found, the new co-reference identification method is at least as effective as a manual search. The number of co-referent matches found however, is higher using the new method. The approach presented here is capable of searching collections stored using differing metadata schemas. More significantly, the approach is capable of identifying potential co-reference matches despite the highly heterogeneous and syntax independent nature of the Gallery, Library, Archive and Museum (GLAM) search space and the photo-history domain in particular. The most significant benefit of the new method is, however, that it requires comparatively little manual intervention. A co-reference search using it has, therefore, significantly lower person

hour requirements than a manually conducted search.

In addition to the overall co-reference identification method, this thesis also presents:

- A novel and computationally lightweight short text semantic similarity metric. This new metric has a significantly higher throughput than the current prominent techniques but a negligible drop in accuracy.
- A novel method for comparing photographic processes in the presence of variable terminology and inaccurate field information. This is the first computational approach to do so.

1

Introduction

GLAMs are a prominent repository of heritage objects and artefacts. As such, they are valuable resources when attempting to understand and study collective and cultural history. One of the primary aims for GLAMs is to make the cultural resources that they hold available to both researchers and the general public for study[8]. In order to assist in this, many GLAM institutions have web accessible collection portals[1]. These portals allow internet users easy access to the institution's records of their collections. Unfortunately, finding particular collection records is often a difficult task[12]. The primary search method offered by collection portals is keyword searching, i.e. a list of search terms is provided by the person searching and any records containing those terms is returned. Keyword based search systems are widely used in many areas due to their simplicity both in terms of use and implementation. Locating desired records in GLAM collections using keyword based searching is, however, difficult for several reasons.

Firstly there is the size of the GLAM collections. A single GLAM collection can easily contain millions of records. The Europeana portal[62], for example, contains more than 21.3 million records from 33 countries and is still growing[178]. The number of distinct GLAMs is also a factor in this. Whilst the essential functionality of the portals remains consistent between collections (i.e. keyword searching), each portal searches a separate set of records. This means that in order to conduct a search of GLAM collection records as a whole, the same search needs to be repeated at each collection portal, which takes time.

Secondly there is the poor quality of the GLAM records. Poor quality in this case does not refer to the record quality as seen by a historian or other GLAM

[1]This is based on an analysis of 18,142 American museums in 2004[8] by the Institute of Museum and Library Services (IMLS). The study found that the majority of museums and large academic libraries made digital records available to the public, the same being true for a lower proportion of public and smaller academic libraries. The World Wide Web (Web) was the predominant method for access.

collections user, but instead as seen by the software searching the records. In particular, GLAM collection records often lack standardised formatting and exist in a large number of different schemas[9, 92][2].

The original motivation for digitising GLAM collection items was conservation[8]. Once digital surrogates of collection items are created, those items are preserved in case of loss and/or degradation of the physical artefact[8, 203]. Given the added collection management benefits that digitisation of collection records brings[3], GLAM institutions have multiple incentives to digitise their collection items and associated metadata. Consequently, digitisation projects have been under way at institutions for decades[4][157, 184].

The rapid increase in internet penetration and consequent uptake in web based services[5], means that internet based resources are an increasingly effective and fiscally effective means of making collections available and accessible. Information can be widely distributed for a relatively low cost. Widespread internet availability has resulted in a changing focus for GLAM institutions. In 2001, the Institute of Museum and Library Services (IMLS) contacted 2,510 museums, at least 40.8% of the respondents[6] replied that one of their main digitisation goals was conservation based[7]. When the survey was repeated three years later, conservation was still a main goal for at least 34.9% of collections[8]. However, increasing collection access had risen as a main goal from 16.9% of respondents to 42.9%[9] [8, 203].

Therefore, many institutions have both collection records which can be made available and, thanks to the change in focus, the motivation to make that information available online. A comprehensive search of GLAM records would involve therefore, multiple institutions, hundreds of separate searches and tens of millions of collection records. Such searches can be required when investigating the output of specific photographers, looking for specific photographs or when looking at overall use patterns of specific photographic processes.

[2]See section 5 on page 80.

[3]E.g. searching, sorting, ease of access.

[4]For example, the American memory project began in 1990.

[5]As of 2011, 77% of UK households had internet access, up 16% in just four years[69].

[6]479 respondents.

[7]Preserve materials of importance or value, 40.8%. Minimise damage to original materials, 35.2%.

[8]Preserve materials, 34.9%. Minimise damage, 12.7%.

[9]Along with the increased access to digital collection, access via the web increased from 23.9% to 36.5%

The change in focus for digitisation projects is also responsible for the poor quality of GLAM records. When the aim was primarily conservation, the records produced could be expected to remain within the originating institution. As such, standardised formats and layouts were unnecessary, and institutions created their own. As a conservation aid, computer readable formats were also optional. What mattered was that the information could be read and understood by future researchers and so standardisation of field contents was also unnecessary.

With the shift in focus towards collection accessibility, collection records are now being made available as a searchable resource and the original non-standardised field formats are showing their drawbacks.

The imprecision of the record information, number of separate collections, variety of collection layouts and size of the collections all combine to make locating specific information very difficult. The limitations of keyword based searching mean that it is barely adequate for the task. Therefore, locating particular information within digital humanities collections is a time consuming task.

1.1 Research focus

This thesis asks if it is possible to use a Computational Intelligence (CI)[10] based approach to assist in searching GLAM collection records and reduce the difficulties experienced by current GLAM collections users. It is hoped that by simplifying searching it will allow collection users to work more efficiently than is possible with the existing systems.

The research conducted for this thesis should be applicable, not just for GLAM collections and records, but also to searching imprecise and uncertain records in general.

In order to provide a focus for this research and a suitable experimental dataset, a particular set of records has been selected as a test case. Therefore, this research will focus on locating matches for the photographic records of the Exhibitions of the

[10]CI refers to a range of approaches inspired by nature which attempt to solve problems which more traditional approaches either fail at or perform poorly against. Examples of CI techniques include Artificial Neural Networks (ANNs), evolutionary computing, swarm intelligence and fuzzy logic.

Royal Photographic Society (ERPS) database[220] hosted by De Montfort University (DMU)[219]. This database contains digitised copies of the Royal Photographic Society (RPS) exhibition catalogues for the period 1870 to 1915. ERPS is one of several photographic history resources hosted by DMU and contain information of 34,197 exhibition entries[11].

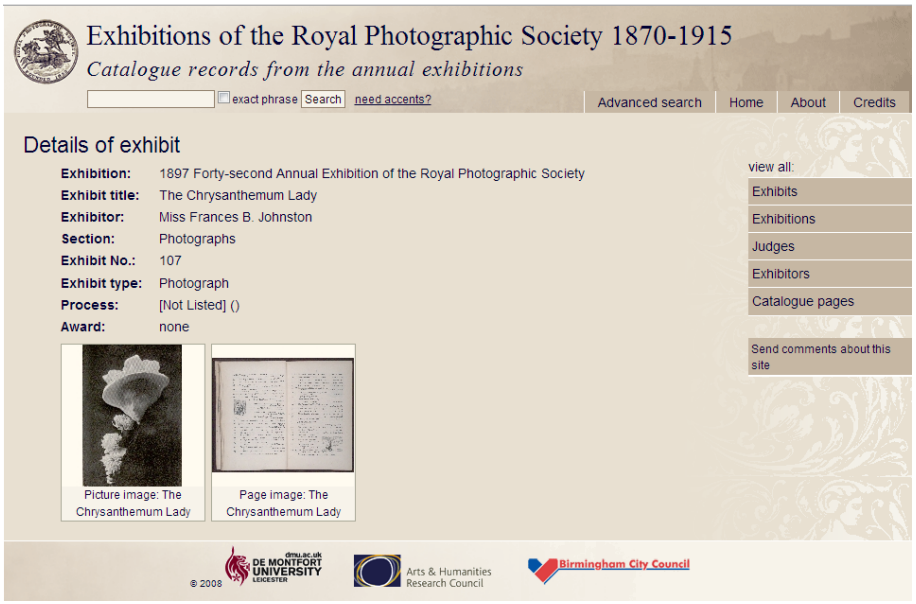
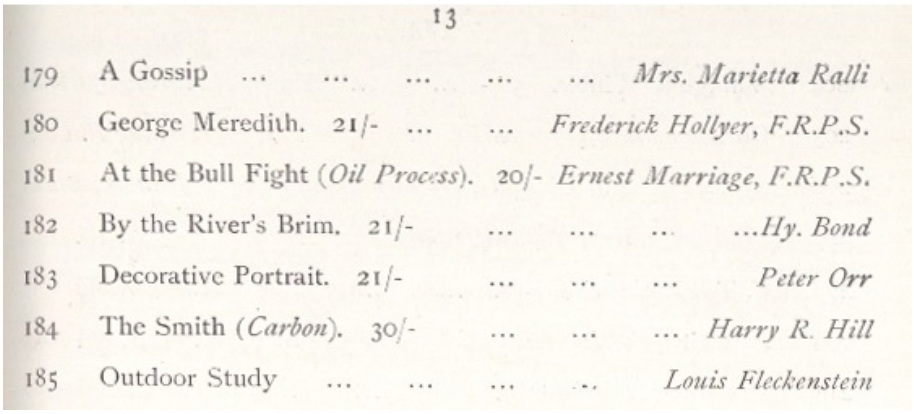


Fig. 1.1: The ERPS collection website showing record erps17654.

As a contemporary account of photography during this significant period of development for photography, the amount of associated information makes the ERPS catalogues unique and consequently valuable for photo-historians. Distinctly absent from most of the exhibition records however, are the actual photographs that the records describe[12]. The reason for this can be traced back to technical limitations when the exhibition catalogues were produced. It was not possible to reproduce all of the exhibited photographs in a time or cost effective way. Consequently only 1,040 of the 34,197[13] exhibit records contain a visual representation of the relevant photograph. The phrase 'visual representation' is used here as not all of the 1,040 records are represented by a copy of the original photograph. Many records are represented by sketches drawn at the time of the exhibitions. Whilst these sketches are better than nothing, they lack the detail of the original photographs as shown by figure 1.3.

[11] See figure 1.1.
[12] See figure 1.2 on the following page.
[13] 3% of the catalogue.



13

179 A Gossip ... ... *Mrs. Marietta Ralli*

180 George Meredith. 21/- ... ... *Frederick Hollyer, F.R.P.S.*

181 At the Bull Fight (*Oil Process*). 20/- *Ernest Marriage, F.R.P.S.*

182 By the River's Brim. 21/- ... ...*Hy. Bond*

183 Decorative Portrait. 21/- ... ... *Peter Orr*

184 The Smith (*Carbon*). 30/- ... ... *Harry R. Hill*

185 Outdoor Study ... ... *Louis Fleckenstein*

Fig. 1.2: Example section from the original ERPS catalogues. Note the lack of images and limited information.

Fig. 1.3: Example of a sketched photograph, erps17094. Note the lack of detail.

As part of its broader aim to improve GLAM collection searching, this research will specifically look at means of locating copies of the 'missing' ERPS photographs in other GLAM collections, the hope being that the ERPS records can be linked to any copies that are found in order to fill in the apparent gaps. Whilst this identification could be achieved by manual searching, the time and resources required to conduct an effective search would be excessive, in part because of the number of ERPS records with missing images. Locating the 'missing' ERPS photographs using CI techniques would demonstrate that automatically locating similar records is possible for GLAM collections.

The focus of this thesis is on records related to photo-history, but it is believed that the methods devised during this research could be used for searching records from other GLAM areas, and potentially applicable for searching in imprecise and

uncertain data sets in general (i.e. domains other than GLAMs).

Identification of similar GLAM records is technically a co-reference identification task[14] and not a generalised search problem[15]. However, information retrieval and co-reference identification are closely related[238, 239]. An approach that identifies co-reference between GLAM records would be directly relevant to a generalised search system for those same records.

Photo-history records were selected as the specific focus for this research for three main reasons. Firstly, a considerable number of photographs appear in in GLAM collections. The associated records are, therefore, well represented in online GLAM catalogues[16] and can act as a good exemplar for the GLAM record search problem in general. Secondly, the 'missing' ERPS images was a known shortcoming for the ERPS website hosted by De Montfort University. There was, therefore, support for an attempt to locate the 'missing' images via any method. Thirdly, De Montfort University has a strong photo-history research centre. The expertise and resources of which would be invaluable in learning about and understanding aspects of photo-history, thereby gaining a better understanding of the search space.

Co-reference identification itself is likely to become increasingly important for GLAM collections. In concert with the increasing amounts of digitised information being made available online[17], has been an increasing recognition that simply creating and storing digital surrogates of collection items is no longer sufficient[42]. Increasingly expected are improved search options and greater interconnectivity between collections[111]. Linked data offers one way to supply this interconnectivity and represents the first step towards the long proposed semantic web[23]. If successfully created, the semantic web would allow for complex and highly detailed querying of web accessible information through the use of software that 'understands'[18] the information that it is searching[23]. Linked data promises to achieve

[14]Location multiple records which both refer to the same real world item, see section 3 on page 34.

[15]Sometimes referred to as information retrieval.

[16]See table 6.1 on page 98.

[17]Sometimes referred to as Open Data[163].

[18]Linked data would not allow software to 'understand' information in the sense of an intelligent, conscious understanding; by explicitly stating connections between pieces of information, it would remove ambiguities surrounding the data and so allow the relatively easy creation of complex search queries and agents which could give the appearance of an intelligent understanding the information.

this by storing information in standard, software understandable formats and by explicitly stating/recording connections between different pieces of information.

By explicitly stating links to information held in other locations, information stored as linked data can leverage those additional collections and make use of additional information when searching. For instance, photographs in collections may have the location that they were taken recorded. Except for very modern images however, this is unlikely to be an exact location (i.e. GPS co-ordinates) but is more likely to be the name of the village/town/county/country in which it was taken. If a searcher is trying to find photographs taken in a certain county, then in a traditional collection they would need to search for every single town, village etc. in that country in order to be sure that they had found every relevant record. In a linked data collection however, location information is just a link to another collection containing geographical information[19]. The geographical collection has all the information on which villages etc. are within which counties. By querying the collections together, a single search can find all the relevant photographs. The photographic collection does not contain or care about geographical information, and the geographical collection does not contain or care about photographs. Each collection can focus on one speciality, and users are able to conduct searches that the collection curators would not have anticipated.

Some benefits to linked data are already becoming apparent as new software programs take advantage of the increasing amount of Resource Description Framework (RDF) formatted information and SPARQL Protocol and Rdf Query Language (SPARQL) endpoints[20] [163].

An example of cross collection searching between GLAM collections with linked data already exists as a prototype produced by Henry and Brown[92]. However, whilst the formats now exist to represent the links between collection records[21], those links are still mainly created by hand. Even the most prominent example of automatically creating links is based upon manual link creation. DBpedia automatically generates RDF tuples based on the content of info boxes in Wikipedia articles. However the info box links were themselves created by hand[14, 30]. Since the links

[19]I.e. Linked GeoData, `http://linkedgeodata.org`.

[20]RDF is the predominant format for linked data information and SPARQL endpoints are the preferred means of making RDF information available[29]. In 2013 there were already more than 63 billion RDF triples and more than 156 SPARQL endpoints in the Linking Open Data (LOD) community project alone[15, 59].

[21]I.e. RDF.

in linked data are created manually, link creation can be a very time and resource expensive process.

A co-reference identification system produced as part of this research could represent the links that it identifies in linked data formats. As it would be able to identify potential links between photographic records automatically, it could prove useful as part of a broader linked data approach for GLAM institutions by reducing the time and resources required to create links.

1.2 The GLAM community's expectations

In order to gain a better understanding of the problem and to provide a guide during the research, a series of informal conversations with individuals in the GLAM community were conducted in order to ascertain their feelings, attitudes and behaviours with regards to digital museum collections both on and offline. As these were informal discussions, they will not be analysed in depth here. However the responses and questions that were received/raised during the discussions were used to generate a series of questions for an online questionnaire which received wider distribution. A call for questionnaire participants was put out to the members of the Museums Computer Group (MCG)[22][81] and the British photographic history website[177]. A total of twenty three individuals responded and completed the questionnaire. The full list of the questions for which can be seen in section I.1 on page 227. The questionnaire was designed to allow for quantitative analysis of closed participant responses. Participants completed the questionnaire individually and online. Although the number of questionnaire responses were limited and more participants could have provided statistical validation to the results, this survey was only intended to act as an initial guide for the research and not a final set of conclusions. The results received provided an initial and early direction which was subsequently refined and changed by the subsequent literature review, personal experience with the search space and discussions with GLAM and photo-history professionals. The results did, however, help to form the initial structure of the final approach with regards to factors such field importance and therefore the fields which were to receive the greatest focus.

The questionnaire was intended to address three main questions, firstly it was to identify what information the responders used and what they considered most

[22] A collection of museum, gallery, archive, higher education professionals and amateurs.

important when searching. Secondly, it determined the number of records that those individuals are willing to, will typically, and would prefer to examine when searching. Thirdly the participants were asked how they feel about recall, precision and general effectiveness of currently deployed search systems in an attempt to determine whether they felt that current search systems were satisfactory for their purposes. A fourth, minor focus was to determine the community's understanding of current search systems. This information would guide which areas of the existing literature and previous research should be focused upon.

1.2.1 Field use/importance

A record consists of several distinct fields, both in GLAM collections and in general. Each field should contain a single piece of information. Different fields will be of interest to the record's reader depending on what they are investigating. In order to identify what information was considered most useful/important when searching, the questionnaire participants were asked what information they used when searching digital collections and which individual piece of distinguishing information they considered the most significant. The questionnaire focused on those fields available in the ERPS records, specifically:

- *Title* - A brief exhibition label for the photograph. This could be descriptive label such as "The Chrysanthemum Lady"[23] or more emotive such as "Solitude"[24].
- *Description* - A longer descriptor of the photograph. This field can contain almost any piece of information including technical details regarding the process used or the location it was taken etc.
- *Person* name - The name of the photograph's exhibitor. This may also be the photographer, but this is not guaranteed.
- Photographic *process* used - The chemical and/or mechanical processes used to create the exhibited photograph and/or the negative of that image.
- *Date* exhibited - The year in which the photograph was exhibited. This may also be the year in which the image was taken in or at least close to it, but again

[23]Taken from erps17654.
[24]Taken from erps20462 and erps24182.

this is not guaranteed. Some photographs appeared in multiple exhibitions.

Some additional options were included when directly mentioned during the informal discussions[25]. A breakdown of the responses can be seen in figure 1.4.

A note regarding formatting within this document. There is a potential for confusion when talking about the record fields as to whether the field itself is being discussed, or the contents of the field, i.e. the *date* field as a whole or just a date within that field. In order to prevent confusion, when discussing a field as whole (be that *title*, *person*, *process* or *date*), the field name will be italicised.

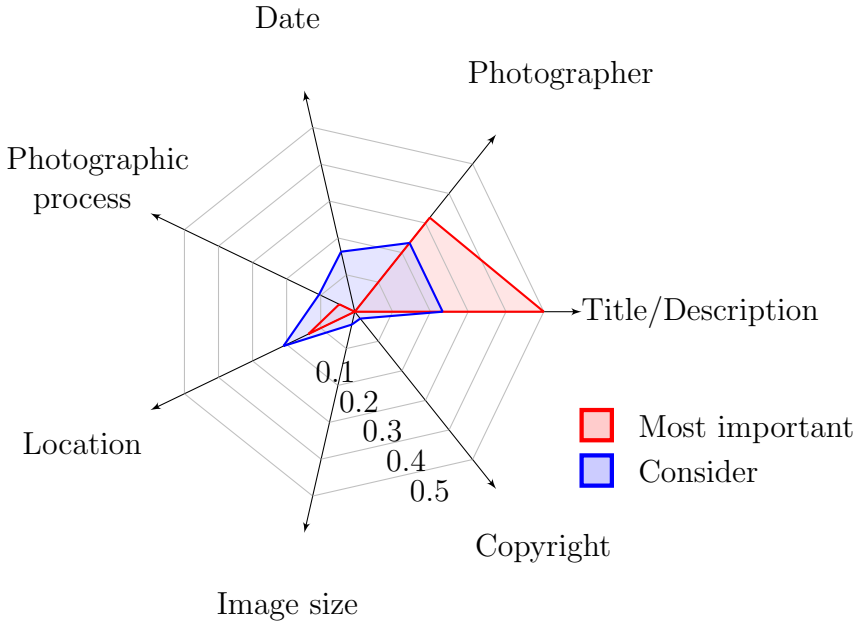


Fig. 1.4: Comparison of field use and importance as reported by twenty three respondents from the GLAM community.

The questionnaire clearly showed that keywords in the *title* and *description* fields were the most widely used, followed by the *person* field. *Date* and *process* were less widely used. Location information for the photographs would be used if it were available for the ERPS records (other GLAM collections do have this information). However, when it came to the importance of those fields, *title* and *description* were the winners, followed closely by *person*. Only a small minority[26] considered *process* and no-one found *date* most significant.

[25] E.g. location, image size and copyright.
[26] 10.5% of respondents.

1.2.2 Number of results expected/desired

An ideal system would return only those results which perfectly match a search query. Realistically the best that can be hoped for is a system that returns a small number of results and where relevant results outnumber irrelevant ones by a large factor.

The questionnaire's participants were asked to provide rough counts for the number of results that they are willing to, would typically, and would prefer to examine when searching. It was anticipated that the users of any system would prefer to examine fewer records than they do at present but that they search through, or would be willing to search through, more in the case of particularly difficult queries.

The results of the questionnaire, as shown in figure 1.5, support this view. At the moment the respondents claim to be examining on average ≥ 80 results per search although they are willing to examine ≥ 122 if necessary. However the average respondent would prefer this number to be brought down to ≈ 58.

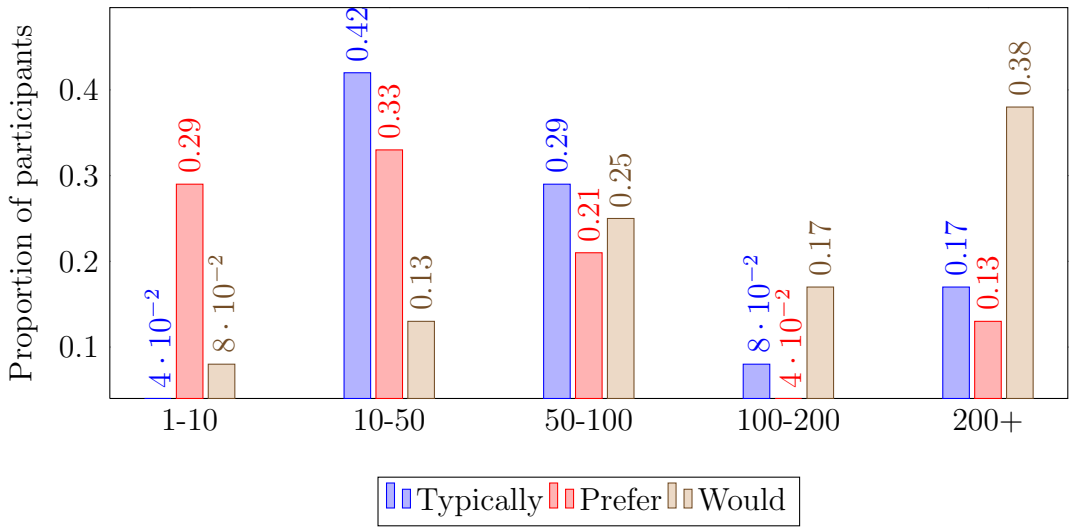


Fig. 1.5: The number of results for a search query that are expected, desired and tolerated As reported by a sample of the GLAM community.

1.2.3 Perceived effectiveness of the current search systems

The purpose of these questions was to determine if GLAM collection users were satisfied with the recall, precision and general suitability of the current keyword based search systems that they use.

The responses that were received show that as far as the GLAM community

attitudes were concerned, the effectiveness of current search systems ranges from bad to acceptable. Only a small minority of individuals viewed the systems positively. This was expected and is not restricted to GLAM collections, user satisfaction in general decreases as domain experience increases[21, 139].However the responses do show that there is a desire for improved search methods.

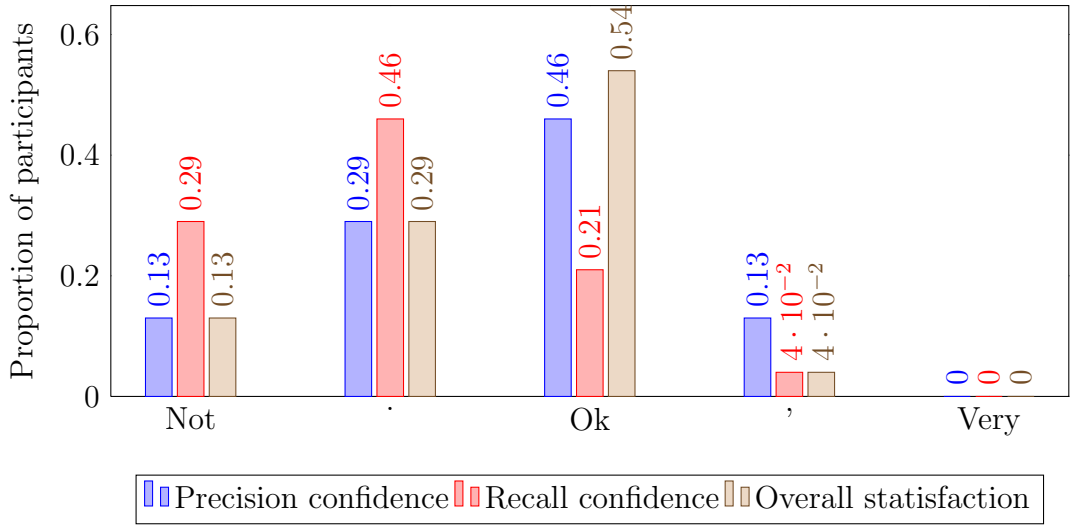


Fig. 1.6: Precision and recall expectations of current search systems. As reported by a sample of the GLAM community.

1.2.4 Search strategy

Beaudoin[21] states that there are two main search techniques. Searchers can start with a narrow, focused query and expand it to make it increasingly generic if the desired results are not found immediately. Alternatively a broad, generic query can be used and made increasingly specific as necessary and based on the returned results. The questionnaire aimed to determine if the GLAM community had a significant preference for one technique over the other. This would have been indicative that one technique produced noticeably better results when manually searching and could therefore have influenced the chosen strategy. However the responses showed that most of the participants use a combination of both techniques depending on circumstance, with a near even split between the participants that only use one or other strategy[27].

[27]See figure 1.7 on the following page.

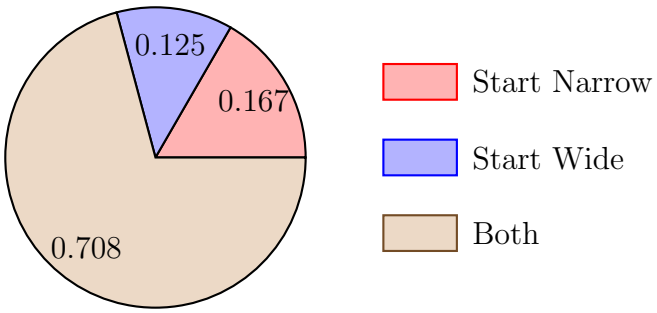


Fig. 1.7: The relative use of different search techniques by a sample of the GLAM community.

Since this research was intended to explore methods for searching across multiple collections, it was felt that it was important to determine the number of collections which were used by members of the GLAM community. The information provides a baseline number of collections to access if human levels of performance and depth are to be approached when searching. A breakdown of the responses to this question can be seen in figure 1.8.

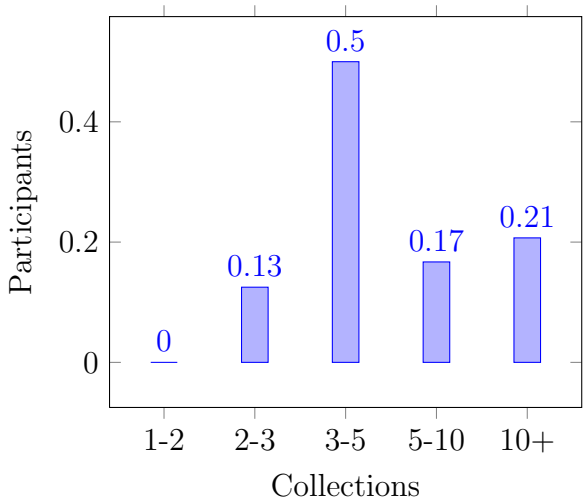


Fig. 1.8: Number of collections regularly accessed by questionnaire participants.

As the responses show, most participants said that they would examine between three and five collections when searching. All of the participants would search at least two and some stated that they would search ten or more. This suggests that although the participants are aware that the records/information that they seek may be found in several locations there are too many collections to search them all. The majority of searchers therefore restrict themselves to a more manageable set of known collections at the expense of better search recall rates.

1.2.5 Conclusion

Whilst keyword search systems are common, the responses from the questionnaire participants demonstrated that the perceived effectiveness in GLAM collections can be classed as merely acceptable at best. The problems with and limitations of keyword searching are well known and referred to repeatedly across the literature[12, 60, 83, 104]. The limitations of keyword searching are, therefore, apparent throughout humanities research and beyond. It is not limited to photographic history. However, one advantage of keyword search systems is that while their performance might be barely adequate, users can easily understand why they get the results that they do from a search query. Search systems which return surprising or unpredictable results may encounter issues in gaining widespread acceptance by GLAM individuals.

There is no preference for either wide or narrow starting searches, with the majority of respondents using a combination of both as the situation demands.

Keywords in the text are the most widely used record feature and were considered the most valuable of search criteria with person information following closely[28]. Whether this is an actual preference by the respondents or simply a requirement given the keyword search based systems that are available is unclear.

1.3 Thesis layout

This chapter has described the broader problems facing GLAMs and the specific case which this thesis intends to address. Also presented were the results and analysis of GLAM community responses to an exploratory questionnaire which were used to guide this research project.

The remainder of this thesis is structured as follows:

Chapter 2 discusses the issues surrounding keyword based search systems and introduces the concepts of query expansion. Included is an overview of both local and global reference approaches with a review of the existing literature related to each. Query expansion and keyword searching are discussed since, for all their flaws, keyword based search systems are the main[29] method of accessing GLAM collection records and so must be dealt with when using GLAM collections and their records.

[28] As shown by figure 1.4.

[29] In some cases the only method.

Chapter 3 covers the basics and set theory of co-reference identification. Also covered as separate sections are descriptions and literature reviews for rule based, Probabilistic Record Linkage (PRL), ANN and cluster based approaches to co-reference identification. Potential pitfalls and issues with the various approaches are identified and discussed. Successfully identifying co-referent records is the fundamental problem facing both this research project and search systems in general. As such the existing and established approaches must be discussed in order to identify which approaches are likely to be successful and which should be avoided during this investigation.

Chapter 4 provides descriptions and analysis of various text comparison/similarity algorithms/methods. Including phonetic, edit-distance and edit-distance resembling approaches. GLAM collections contain large amounts of text. As such, searching GLAM records requires that the search approach can handle the problems of textual information[30]. Text similarity algorithms will be necessary in any search system created and so the existing and established approaches which may be applicable are discussed.

Chapter 5 presents a discussion of the formatting, structure and access to GLAM collections. This chapter also discusses the major issue with GLAM collection records along with the reasons that caused it.

Chapter 6 presents the significant contributions of this thesis. The full sequence of actions necessary to search for co-reference between a single record and multiple GLAM collections is described. The individual field similarity metrics/algorithms are discussed along with the reasoning behind their designs, worked examples and algorithms when appropriate. Also included are details regarding failed approaches when said failures directly influenced or led to the approaches which were finally produced.

Chapter 7 details the testing of the approach and similarity metrics described in section 6. This includes the reasoning behind and explanation of the testing methodology used, problems that were encountered, the final results and analysis

[30]Described in detail in the chapter.

of those results.

Chapter 8 contains a final summary of the research. This includes identifying areas of significance, whether the research met its intended targets, a final analysis of the testing results and presents potential areas and directions for future research.

2

Query expansion

Having established the aim and scope of this research, this chapter now turns to an analysis of keyword based search methods, the primary problem with them, the dominant technique for mitigating that problem (i.e. query expansion) and the various query expansion approaches available.

The vast majority of search systems for both GLAM collections and in general function on the basis of keyword searching. That is to say that they work by simply comparing a list of provided terms against every one of the records in the collections being searched[1]. Matches are identified as those records containing one or more of the search terms[93]. Those items where the terms do appear are then returned as results to the user. Keyword based methods are simple to implement, easy to understand, often effective and widespread. Keyword search has been a fundamental part of computer systems since the earliest days of the technology as it is just an extension of a simple find function. Although the earliest instance of keyword searching is unclear, evidence of its use dates back as far as 1948 and the UNIVersal Automatic Computer (UNIVAC) machine[189]. The systems operate, however, on the fundamental assumptions that one or more of the terms being searched for must appear in the text of a record in order for it to be included in the search results. This means that selecting the correct words to search for is the most important factor in using a keyword search system.

The results of a search query can be judged according to two criteria, precision and recall. Precision refers to the number of results which are relevant as a proportion of the number of results returned. Recall describes the number of relevant results returned as a proportion of all the relevant results which could have been

[1]Whether the terms are compared against every field in the records or a subset depends on the configuration and nature of the records and search system.

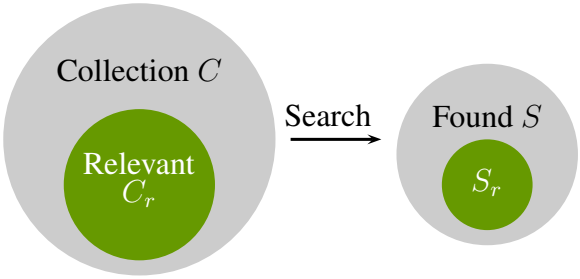


Fig. 2.1: Visual demonstration of precision and recall in searching.

returned (i.e. all the relevant records in the search space). If C is defined as the full collection of items being searched and S as the results selected from C by a search query (see fig. 2.1), then $S \subset C$, and the recall (R) of a search query can be described as $R = S_r/C_r$. If r is further defined as only those items which are deemed relevant to the search query, then precision (P) can be described as $P = S_r/S$. The overall success of a particular search can be calculated from the precision and recall values. These are combined to generate a measure of the query's accuracy known as a F score[2]. This is calculated as shown in equation 2.3[26], assuming that equal importance is given to each factor.

$$R = \frac{S_r}{C_r} \tag{2.1}$$

$$P = \frac{S_r}{S} \tag{2.2}$$

$$F = \frac{RP}{\frac{1}{2}(R+P)} \tag{2.3}$$

The preferred outcome is to achieve both high precision and high recall, ensuring that only relevant results will be returned. Realistically however, increases in precision will typically produce a decrease in recall and vice versa[32, 77].

Whilst there is some cross-over with methods for improving search precision, query expansion is mainly focused on improving the recall of search methods by augmenting the terms of a search query. Thanks to the flexibility of most languages, most items can be described in several different ways using different terms. This distinction between an item and its lexical description is the cause of the main problem with keyword search methods and is referred to as synonymy. If the terms

[2]Also called F_1 score or F measure.

used for the search do not match the terms used to describe the item then despite the item matching the search meaning, the lexical difference means that it will not be chosen. For example, a photograph of "a flock of geese" will not be found if the search terms used are "goose" or "birds" though these are valid descriptors for that photograph.

Since the only factor excluding these semantically similar items from the search results is the absence of the correct keywords, solutions for this problem have focused on adding additional terms to the initial search terms in order to lexically encompass all relevant items. This augmentation of the original search query can be achieved using several different methods, and a combination of the techniques can be employed in order to achieve the maximum benefit. These methods are:

1. Spelling Correction - Typographical and spelling errors in both the search terms or the searched records will negatively influence the number of results returned from a keyword search since the misspelt term will only match against those items which misspell the same word in an identical way.

2. Stemming - A single word will have various forms depending on the circumstances in which the term is being used. For example, an individual noun will have both singular and plural forms whilst verbs can have different past, present and future forms each with different first, second and third person variations. For example, the verb "to swim" has the additional forms of "swimming", "swam", "swum" and "swims".

3. Synonym expansion - The inclusion of semantically similar but lexically different terms. 'Goose' for example has synonyms of 'gander', 'bird', 'waterfowl' etc.

2.1 Spelling correction

Given that the records of interest for this research are from GLAM collections, the problem of spelling and typographic errors is not as great as would be faced by a more generalised search system (i.e. internet search engines) where the items being searched can be and are created by anyone. GLAM collections are typically created and maintained by GLAM institutions and the professional curators working there[3]. Although a great deal of care and attention is given to collections, since the record

[3]Exceptions do exist, particularly in smaller institutions where volunteers may be used instead.

information will have been manually created at some point it is a near certainty that some errors will exist.

Methods for spelling correction can also be utilised for spelling normalisation. English has a number of regional forms which can influence the spelling of many words. The most obvious of these regional variations are the differences between British and American English which can effect the spelling of individual words (i.e. 'colour' and 'color') but also entire series of words through differing suffix styles etc (i.e. -ise and -ize).

When automatically dealing with typo/spelling errors the standard strategy is to use a whitelist of all accepted words[140]. Any words which do not match this list are then either not accepted or are altered to match the nearest term in the list. Whilst undoubtedly efficient at preventing spelling mistakes, simply blocking unknown words is only acceptable for very limited vocabularies. Regardless of the whitelist used, it is likely there will be valid terms missing from it. The inability of this method to handle uncommon person and place names is a severe limitation.

Despite these flaws the whitelisted approach is the dominant technique[146]. The difference between different implementations comes down to the process used to identify which whitelisted word most closely resembles any unrecognised terms and the word list used. Identifying the word which was intended necessitates a process of measuring the 'closeness' in spelling of distinct words. A large number of techniques exist to measure this 'closeness', though these will not be discussed in this section. For details of various textual similarity measures and comparisons of such methods see section 4.1.

Spelling correction can not be used to its full potential in this project. Although it would be possible to correct typographical errors in the ERPS records, it is not possible to fix the external GLAM collections. This means that if a word was spelt incorrectly in an external collection, records containing it would not be found when searching using the correct spelling. This does not mean that spelling correction should be discounted entirely. Correcting errors in the ERPS records should still seen as worthwhile given the international nature of the internet and the GLAM collections which it makes accessible. Including regional variations on a word's spelling in the query expanded search terms should improve the recall rate when examining GLAM collections from other countries.

2.2 Stemming/term expansion

Almost every word in the English language has various forms in order to distinguish the tense or an action or the number of items being described. These inflected forms can dramatically affect the spelling of the source word (i.e. 'goose' and 'geese', 'swim' and 'swimming') and so the use of the incorrect form with a keyword search system can severely reduce the recall of a query. This recall problem can be resolved if the alternate inflected forms of a search keyword are included in the search query at the expense of the precision of the query.

In order to identify a word's inflected forms it is first necessary to reduce the word to its root or stem form, this process is called stemming. This process is can be achieved using one of several established methods including the Dawson[50], Krovetz[113], Lovins[129], Paice/Husk[165] and Porter[175] stemmers. These are all rule based methods, in which a suffix rule list is consulted and the appropriate substitution made. The rule list varies between the techniques but many of the rules remain constant across multiple methods, Dawson for example is a direct descendent of Lovins and uses a modified version of its rule list (expanded to $\approx$ 1,200 rules). As multiple rules can be triggered by a single word, several approaches (Dawson, Lovins) select between rules by identifying the longest suffix match in the rule base.

The performance gains from including various inflected forms varies by the language being stemmed but when processing English resources the performance gains are generally considered to be significant[113]. However, searching items which have had their searchable text reduced to stem forms produces variable results with some searches producing better and other searches producing worse outcomes to achieve no overall benefit[86]. A comparison between several stemming approaches likewise notes no significant difference in performances[97, 98].

Once the word is in its source form the most accurate method for stemming is to use a lookup table to identify the different inflected forms. This strategy will of course fail if the word to be stemmed does not appear in the lookup table. This can occur if the word in question is a recent invention or is relatively rare. Technical and domain specific terminology suffer from this problem the most.

Alternatively an algorithmic approach can be used to generate the inflected forms for a given term using grammatical rules. Given the patchwork nature of the En-

glish language however, there is no simple set of rules which will generate correctly inflected words from a given root form. If the term in question is unusual or does not follow the typical rules then an algorithmic approach can produce incorrect inflected forms (i.e. 'geese' not 'gooses' is the plural form of 'goose').

Algorithmic methods are accurate for the majority of words and the process can be hybridised with a lookup table in order to handle prominent exceptions to the algorithm's rules.

For the purposes of query expansion it can be effective to include, alongside the search terms, the terms' various inflected forms[103]. Even when incorrectly inflected forms are included this does not negatively affect either the precision or recall of a search to a significant degree. This is because the incorrect forms are unlikely to match against the items being searched.

As with spelling correction, converting the terms in the records of the external GLAM collections to a single standard stem form is not possible. The inclusion of multiple inflected forms in the query expanded search terms should, however, produce a noticeable improvement in the recall rates of the search queries.

2.3 Synonym expansion

Synonym expansion[4] is the most challenging of the query expansion techniques. Synonym expansion aims to add semantically related but lexically distinct terms to the search query. Perfect synonym expansion requires an understanding of the meaning of the search query[18]. As this is generally considered to be an Artificial Intelligence (AI) complete problem[18, 119][5] it is not achievable at present. Therefore, whilst synonym expansion will increase record recall[137] it can also significantly decrease query precision. This is an example of the so called precision-recall trade-off[32, 77, 80]. For example record 17093 from the ERPS collection[6] is titled "Fair Daffodils". Synonym expansion of 'daffodils' can generate terms such as 'flower' and 'bloom' which are valid alternatives. However, expansion of 'fair' can produce not just 'attractive' and 'beautiful' but also 'impartial' and 'carnival' which are not suitable given the context. Despite this problem, current techniques produce suffi-

[4]Often referred to in the literature as just "query expansion"[137].

[5]I.e. that solving it would require solving the central AI problem, how to make a computer which is as intelligent as a person?

[6]Referred to as erps17093 from here on.

ciently accurate results for synonym expansion to be a valuable addition to keyword based search systems.

2.3.1 Global reference approach

Synonym expansion can be easily divided into two major approaches, global reference and relevance feedback[137]. Global reference query expansion identifies synonyms for a given word by means of a lookup file. In effect this process can be described as the automatic use of a digitized thesaurus. Despite the apparently simple nature this method is effective.

The central concern for the use of global reference based methods is the origins of the synonym lookup file. Whilst simple flat files of known terms and their associated synonyms can be and are used (i.e. Roget's Thesaurus), also used are so called Lexical DataBases (LDBs) are also used. LDBs contain significantly more information than just a list of terms and synonyms though they do also provide that essential functionality[145]. The best known and most widely used LDB is WordNet[66, 144] which includes 155,000 words stored in 117,00 synsets as of WordNet 3.1[223]. A synset is a collection of related and potentially interchangeable synonyms. Each WordNet synset is linked to a small number of other semantically related synsets, most of these links follow a hierarchical IS-A[7] structure identifying hypernym and hyponym relationships between the synsets[144, 145]. For instance 'flower' is a hypernym of 'chrysanthemum' and is in turn a hyponym of 'plant'. WordNet also includes links representing meronyms[8], holonyms[9] and antonyms[10] as well as word definitions and examples of word usage.

Creating synonym lookup resources (either flat files or LDBs) is traditionally a very time and resource intensive process[194]. This is especially true for LDBs due to the time needed to create the internal semantic links. Fortunately there are freely available resources already available (i.e. WordNet[66] and Roget's Thesaurus[110]) but these are designed as general purpose references and so lack domain specific terminology or a specific domain focus when it comes to acceptable synonyms. When possible, use of a domain specific corpus is considered preferable[100] as the

[7] X is a Y.
[8] X is part of Y.
[9] X has a part Y.
[10] Opposite of a synonym.

chances of incorrect synonyms being included in the search terms is reduced. The number of domains with specialised corpora already available is, however, limited. For instance the Medical Subject Headings (MeSH)[159] controlled vocabulary maintained by the U.S. National Library of Medicine is intended for use in indexing articles from medical journals but can be and is also used as a source of synonyms for global reference query expansion.

Combining a large, existing, general purpose LDB with a smaller domain specific one can reduce the problems posed by domain specific terminology whilst still benefiting from most of the time and resource savings available from using a pre-existing LDB[131, 133, 198]. Whilst this method does solve the issue of a lack of domain specific terminology, this approach still suffers from a lack of domain focus on those terms contained in the general purpose collection.

2.3.1.0.1 Automatic LDB production Due to the significant time and resource requirements for manual thesauri creation, methods for automating this process have been the focus of many research projects[91, 126, 162, 226].

There are two main methods for automatic production and regardless of the method used, automatic generation for synonym lookup resources requires that a large corpus of text is available for analysis. The created thesaurus will then be specific to the domain of that corpus of text, this means that if the new thesauri is intended for use in an area with a large amount of domain specific terminology it is vital that the text corpus analysed is representative of that domain area. General purpose resources need to analyse a comprehensive range of text from a wide variety of sources in order to avoid becoming overly specialised in one area.

Unless there is a pressing need otherwise, the generated resource will benefit from manual corrections and cleaning. Automatic generation relies on automatic identification of related terms, this can be achieved using the two approaches described below:

Statistical term proximity Term proximity, also called co-occurrence identifies pairs of words which appear in close proximity a statistically significant proportion of the time. A common approach is for each record[11] in the corpus to be

[11]Record in this case generally means a document but could refer to any piece of text.

represented as a term vector. Each document represents a single dimension in n-dimensional space where n is the number of documents in the corpus being analysed. An example set of documents and the term vectors which would be derived from those documents are shown in tables 2.1a and 2.1b on the next page. For the sake for simplifying the model, some low value words[12] have been excluded as they most likely would be in a real system. Potential synonyms can then be identified by identifying the column vectors of the term vector matrix which closely resemble each other.

The cosine similarity measure is the most commonly used method for calculating vectorial similarities. Given term vectors a and b this method simply finds the cosine of the angle between the two vectors[13]. Since only the relative angle of the vectors is considered, this means that the cosine similarity method functions on the basis of the proportions of terms in the vectors and not the relative magnitudes of the vectors. In practise this means that differences in the length of the documents and, therefore, the number of times that each term appears does not directly effect the results.

In their 1992 paper however, Chen and Lynch describe an asymmetric method which they refer to as the Cluster algorithm[14][35].

$$cluster(a, b) = \frac{a \cdot b}{a} \tag{2.4}$$

$$cluster(b, a) = \frac{a \cdot b}{b} \tag{2.5}$$

For both algorithms the pair-wise similarity for all term vectors is then calculated and the pairs with the highest similarity values are chosen. Since the cluster approach is asymmetric it requires double the number of pair-wise similarity value be calculated compared to the cosine method. The cosine algorithm can simply mirror the similarity values since $cosine(a, b) = cosine(b, a)$, this allows for a full set of t^2 pair-wise values (where t is the number of term vectors) to be simulated using only $\frac{1}{2}(t^2 + t)$ values. This effectively halves the number of calculations required and given the number of terms likely to be contained in a corpus this represents a significant time saving. Chen and Lynch state however, that the Cluster algorithm has

[12] I.e. 'the', 'a', 'and', 'is'.
[13] See section 4.2.3 for a full description of cosine similarity.
[14] See equation 2.4.

a higher concept recall rate[15] than either cosine or a manual approach but that the concept precision rate[16] for both methods remained similar and inferior manually produced lists of synonyms[35, 192]. This means that the Chen and Lynch method selects more terms but that an equal percentage of the terms should not have been chosen. It is not clear if the increased expansion performance of the cluster approach justifies the increased computational requirements.

Document	Text
1	A duck is a bird
2	Look at that bird, is it a duck?
3	Crispy duck meal
4	A romantic meal

(a) Example text documents.

	bird	crispy	duck	look	meal	romantic
1	1		1			
2	1		1	1		
3		1	1		1	
4					1	1

(b) Term vectors for example documents.

	bird	crispy	duck	look	meal	romantic
bird	1.0		0.82	0.71		
crispy		1.0	0.58		0.71	
duck			1.0	0.58	0.41	
look				1.0		
meal					1.0	0.71
romantic						1.0

(c) Cosine similarity of column vectors shown in table 2.1b.

Table 2.1: Example use of cosine similarity on a corpus for identifying synonyms.

Lexical analysis Whilst this method is repeatedly referred to in the literature there is limited evidence of this strategy being deployed in the real world. Based on the prevalence of literature on statistical techniques, lexical analysis appears to be the less widely used method. There are however, still some well documented implementations such as the TINA and Context OPerator SYntax (COPSY) projects from the Siemens Natural Language Processing (NLP) group[191, 196], the hyponym

[15]The number of relevant terms selected
[16]The number of terms selected that were relevant

identification approach by Hearst[91] and Semantic EXtraction from Text via Analyzed Networks of Terms (SEXTANT)[79].

SEXTANT was developed and described by Gregory and consists of four main stages[79, 192]. These are:

1. Lexical analysis - the terms in the text are identified and separated (tokenisation of the text). The tokenised terms are then processed to identify the most what part of the text they most likely represent (i.e. nouns, verbs, adjectives etc.).

2. Bracketing - Phrases are extracted from the text using a rule based method. For instance adjectives can be used to modify nouns so occurrences of a noun following an adjective, those two words would be bracketed as a possible phrase. Using a series of rules common sentence structures can be identified and extracted.

3. Parsing - Syntactical relationships are then extracted from the bracketed phrases, this identifies the contexts for individual terms based upon the other terms in the same brackets. Gregory achieves a 75% accuracy rate using a five pass process. This could be increased either by using better parsers or more passes[192] but this would increase the computational cost for limited gains.

4. Term similarity - Under SEXTANT only the similarities between nouns are calculated, the remaining terms are ignored. Similarity is calculated using a weighted Jaccard metric[17]. The aim is to discover words which are used in similar ways, that is to say that the brackets that the nouns belong to resemble each other. The nouns from similar but different sets of brackets can then be recognised as being related to each other.

As with any LDB creation technique, the related terms which are identified are highly dependent on the corpus used during the creation. For example the terms associated with the word 'case' can be dramatically different depending on the text which is analysed. An examination of medical text would identify terms such as 'patient', 'disease' and 'treatment. An analysis of articles related to Kennedy assassination conspiracy theories however, would identify terms such as 'evidence', 'investigation' and 'conspiracy'[192].

[17]See section 4.2.2 on page 71.

Web based approaches Whilst statistical and lexical analysis techniques may be predominate, an alternative and interesting approach has been proposed by Gabrilovich and Markovitch[74]. Their paper points out that the relatively recent emergence of a freely available online encyclopaedia in the form of Wikipedia[70] provides access to a large corpus of both ordinary text and domain specific terminology. Importantly the links between related concepts are already present and, having been manually created, should be relatively accurate. Since the formatting of Wikipedia articles is comparatively standardised when compared to other sources, this means that mining the articles (and the links between them) is simple. In their work Gabrilovich and Markovitch use the anchor text[18] of Wikipedia articles to identify related concepts. This method places the understanding of the domain fields with the Wikipedia contributors and so can cover a very large range of domain areas. Whilst the level of domain specific terminology falls short of custom resources (e.g. MeSH) it is an improvement on generalised resources (e.g. WordNet).

In many ways Wikipedia is ideal for this purpose. The time and person power resources invested in it[19] exceed any potential academic or commercial venture. As an added benefit the collaborative and continuous revision of the site's articles ensures that the synonym terms identified should remain up to date. The work by Gabrilovich and Markovitch is not the only method to utilise Wikipedia, as evidenced in the work of Strube and Ponzetto[205].

The use of Wikipedia however, only represents some of the most recent attempts to leverage the vast amounts of text available on the Web for synonym identification. For example Turney presented a method of calculating a similarity value for any pair of words using search engine results in 2001[192, 218]. The method uses a set of four search queries that make use of the boolean search options available in some keyword based search systems[20]. The number of results returned for each search can then be used and combined to confirm or show the likelihood that the two words are related. Given that this method produces a similarity likelihood values for term pairs, rather than a list of synonyms, it can not realistically be used to create LDBs resources on its own. It would be necessary to compare every[21] word in a language to every other word which would take a long time. It can however, be used to validate the results from another synonym identification system with

[18]The visible text of a Web hyperlink.

[19]22.8 million articles by 1.5 million users as of May 2012[71].

[20]The AND, OR and NOT operators, also used is a NEAR operator but that is not a standard operator.

[21]Or a significant proportion.

a high degree of accuracy, using as it does, the vast amount of multi-domain text available on the Web.

There are therefore, multiple methods for generating a LDB based on GLAM records. The need for a custom LDB as part of this research however, appears unlikely. Custom LDBs are only worth the extra time and effort they need when they focus of a particular domain or subject area or when the text being examined contains a significant amount of technical terminology. Whilst the records being examined for this research do include some technical terminology[22], the majority of the text consists of ordinary language descriptions of the contents of photographs. As such, there is no particular domain focus to be explored since the photographs cover a comprehensive range of subjects. Therefore, if LDBs are used in this research then the pre-existing, generic ones are expected to be sufficient (e.g. WordNet).

If generic LDBs do prove to be insufficient then the use of lexical analysis techniques to generate a custom LDB seems unlikely. The sentence structure used in GLAM records often does not consist of full sentences[23]. There is, therefore, a concern that the rules used by existing lexical analysis techniques would perform poorly given the truncated nature of the text. Statistical analysis techniques are expected to perform better in this situation.

2.3.2 Relevance feedback approach

Relevance feedback[24] is the second significant approach for synonym expansion in query expansion methods. An iterative process, relevance feedback uses the text of the results returned in each iteration to identify additional search terms. A search is performed using an initial set of search terms and the results retrieved. An examination of the results is conducted and the relevant (to the search query) results identified manually. These relevant results are then analysed to identify new terms which are not currently included in the search query. The precise selection criteria for new terms varies, but is typically based on the number of occurrences of the new terms or the physical proximity of the new terms to the existing search terms in the results' text.

[22] Photographic processes etc.
[23] In particular in the *title* field.
[24] Also called local relevance feedback.

2.3.2.1 Pseudo-relevance feedback

Since relevance feedback requires manual interaction with the system in order to distinguish the relevant results, its usefulness in fully automated search methods is limited. A variation on this method has the results ranked and ordered according to their algorithmically determined relevance. The actual relevance of the top k results is then simply assumed[17]. Exactly how the results are ranked and consequently the top k results are selected depends on the records being processed but some kind of similarity metric for the records is required. Standard metrics for records containing text are cosine similarity, Okapi BM25 and BM25F in the case of records with more than one textual field. These techniques are discussed in greater depth in section 4.2.

The work by Harman [87] demonstrates that both pure relevance and pseudo relevance are effective at identifying additional results but that pseudo-relevance requires a larger number of iterations in order to identify said results. This is not unexpected given that manually selecting the relevant results would be expected to produce more accurate results than an automated method but is in any case irrelevant. Pseudo-relevance feedback approaches have proven effective in certain situations[11, 115, 214] and algorithmically ordering the results is so much faster than manual identifications that, even with the extra iterations, the pseudo-relevance method is quicker. This speed comes however, at the cost of an increased risk of topic drift[87] and inferior results in a direct comparison to 'real' relevance feedback.

2.3.3 Comparison of synonym expansion approaches

Both global reference and relevance feedback techniques have their advantages and disadvantages. From a conceptual and implementation standpoint, global reference is the simpler technique but the need for a LDB or similar raises difficulties.

Relevance feedback does not need an LDB at all. It uses the text contained in the results to identify new terms, this also bypasses the issues of domain focus/terminology. However, relevance feedback applications in the literature operate on search spaces which have significantly more text available per item than is seen in records taken from GLAM collections. Local reference is often used in document search/classification systems where each item can contain hundreds or thousands of words. The text available from the collection records may be so brief that feedback methods will be unusable.

2.3.3.1 Topic drift

The significant concern for both synonym expansion approaches is that neither understands the meaning of the text being expanded. Many words can have several distinct meanings[25] and even a single meaning can take on different semantic inflections depending on the context[26]. This means that even though the extra terms identified may be valid synonyms for the current search terms when taken individually, it does not mean that it is valid within the overall meaning of the query. The inclusion of unsuitable terms can cause an issue known as topic drift. Topic drift manifests as the returning of records from another unwanted domain due to one or more of the search terms also/only being a valid term in the unwanted domain. For example record erps17093 is titled "fair daffodils" and although the meaning of 'fair' in the context is 'light', 'blonde' and/or 'beautiful' another valid meaning for the word (though not in this context) is "a gathering of stalls and amusements for public entertainment"[197]. If synonyms for this second meaning of the word are included (i.e. 'market', 'fete'), then the focus of the search will be affected and irrelevant results returned. As relevance feedback is an iterative process, the search query terms can drift further and further from the original focus as more and more terms are included on the basis of irrelevant results.

For global reference the inclusion of non-relevant terms only means that some irrelevant results may be included. The non-iterative nature of a global reference approach means there is no reinforcement of the incorrect concepts. Topic drift is a problem for pseudo-relevance feedback rather than relevance feedback as a whole. When a person selects the relevant results for use in the expansion the chances of irrelevant results being included is greatly reduced. Topic drift can still occur but it will be spotted and the query either re-run or the results invalidated.

There are a number of methods which can be used to mitigate and/or limit the influence of irrelevant words once they have been included in the search query. These will be discussed later in sections 2.3.3.2 and 2.3.3.3. The ideal solution, however, is to prevent the inclusion of these irrelevant terms in the first place. This means identifying the semantic meaning (word sense) of the search terms in the particular context of the search query be identified. Unfortunately Word Sense Disambiguation

[25] Homonyms.
[26] Polysemes.

(WSD) is an incredibly difficult problem to solve and is generally assumed to be an AI-complete challenge[152].

WSD has been a significant problem for NLP and various methods have been tried with varying levels of success. What is mentioned[227] as a significant stumbling block to successful WSD is insufficient quantities of text. Current approaches[134, 227] require a series of related terms to appear in the text in order to identify word sense. Given the briefness of the text in most GLAM records, successful WSD would be difficult.

2.3.3.2 Number of terms

With both methods of query expansion consideration needs to be given to the number of additional terms which will be added to the original search query. With global reference a hard limit is enforced by the number of synonyms which can be found for the original terms but with reference feedback it is theoretically possible to add every word in the English language given enough results and enough iterations. Such a set of search terms would be completely useless but more realistically, relevance feedback can easily add hundreds of additional search terms.

The literature existing is widely divided on this issue, ranging between 20 terms[185], a third of the number of the original search terms[87] to colossal expansion in the region of 300 - 500 additional terms[33]. Those implementations which use a relatively small number of additional terms appear to place greater emphasis on selection of quality terms whilst those implementations using extensive expansion rely on the sheer number of relevant terms to mitigate the effect of a few irrelevant additions.

2.3.3.3 Additional term weighting

Any automatic process for selecting additional terms will result in the inclusion of some irrelevant ones. As such the level of faith/trust which is placed in the new terms should be of concern. Whilst it can be safely assumed that all the terms in the original query are highly relevant to the desired results, the same can not be said of the additional terms. Some query expansion methods handle this reduced level of trust with a term weighting value which places lower importance on the new terms. Other implementations take the opposite approach and apply a greater weight to the additional terms, applying the greatest weight to those terms with the lowest occurrences.

Whilst weighting the additional terms will be considered for this research, actually implementing it may prove challenging given the limited access available to the external collections.

2.4 Conclusions

This chapter described the dominant search method used, both in digital GLAM collections and in digital information searching in general(i.e. keyword based searching)[190, 195, 235]. Also discussed was the primary problem presented by that method (i.e. that exact matches between the collection records and the search keywords are required). If not addressed, the known problems of keyword based search system would cause low recall rates[27] for any automatic or semi-automatic search system created during this research. Whilst the use of spelling correction and stemming for this project may be limited, due to the limited availability of the records to be searched, synonym expansion seems likely to play a large role. By expanding the original terms contained in the records being searched for to include semantically similar terms and the various inflected forms of those terms, synonym expansion (in combination with stemming etc.) looks like a promising way to ensure that acceptable recall rates are achieved for any eventual search system. The widespread literature on the use of query expansion methods in document classification systems strongly suggests that this is an effective technique, though its effectiveness on GLAM records has yet to be proven and the briefness of GLAM record text means that relevance feedback methods will need to be avoided[28].

[27] The proportion of valid results in the search space as a whole which are successfully returned in the search results, see section 2 on page 17.

[28] See section 2.3.2 on page 29.

3

Co-reference identification

The previous chapter discussed the problems posed by the existing keyword search interfaces and the difficulty in getting the GLAM collection interfaces to return their relevant items. When relevant items are returned however, they will be amongst a number of irrelevant ones[1], the proportion of relevant to irrelevant describes the precision of that search[2]. This chapter describes various techniques for separating the relevant and irrelevant items.

In order to locate copies of the missing images from ERPS, the ERPS records need to be compared against the records of other GLAM collections. By finding records in GLAM collections which closely resemble those in ERPS, it is hoped that it will be possible to find multiple records referring to the same photographs. If one of the matching records found in an external GLAM collection has a copy of the photograph it is describing, then this could be used linked to by the ERPS records. In other words, the need is to identify when two distinct records in two distinct locations represent the same real world item, in this case a photograph. Finding and identifying similar records is a process variously referred to in the literature as co-reference/record/entity identification/resolution/linkage/matching[37]. However, for the purposes of this review the terms co-reference identification and linking will be used.

Ideally linking records would be done using Unique IDentifiers (UIDs). If the linkage is being conducted between multiple collections, then the identifier would

[1] I.e. items which are not of interest to the searcher, mistakes in the search process etc.
[2] See section 2 on page 17.

need to be common to all the collections[3]. Examples would be International Standard Book Number (ISBN) numbers or National Insurance (NI) numbers. Barring the inevitable occasional errors in the data, unique identifiers offer the fastest, easiest and most accurate way (near 100%) to perform co-reference identification. Co-reference identification becomes significantly more challenging when the records being linked are either missing their unique identifier or when no UID exists. For example, all books have an ISBN which uniquely identifies a specific work[4]. If all of the records referencing a specific book have the ISBN numbers recorded, then it is trivial to see if the records are co-referent. If the ISBN numbers are missing then it is necessary to resort to using combinations of multiple pieces of identifying, but not uniquely identifying information. In the case of books, a comparison of the title and the author would generally be sufficient. Whilst an author may have multiple books and there may be multiple books with the same title, an author is unlikely to have given the same title to multiple works. Therefore by combining fields, specific works can be identified and probable co-reference can be established in the absence on a UID.

Since records may be collected from multiple different sources, the same information may be represented in a number of different formats (unlike UIDs). Therefore, good co-reference systems display intelligent, human-like behaviours by identifying the underlying similarities between the individual features of the records or converting the features into a single standard representation before combining multiple features in order to identify matches.

Whilst co-reference identification as a concept can be traced back to 1946[57], the formal mathematical basis and underpinnings are normally attributed to Fellegi and Sunter. In their original paper[67], they describe record linkage as two sets of items A and B whose elements are defined a and b. Some elements are common to both A and B, and these are the co-reference items are of interest. Record linkage is, therefore, the process of identifying those common elements.

The set of ordered pairs of the elements from A and B (equation 3.1) can, therefore, be split into two disjoint sets of the matched (M, see equation 3.2) and unmatched (U, equation elements 3.3) [38].

[3]A Globally Unique IDentifier (GUID). There is a difference between identifiers that are unique within the confines of a single collection/database etc. and identifiers which are used across multiple collections/databases. Within this thesis UID will be term normally used unless the point is specifically referring to cross collection/database etc. identifiers.

[4]ISBN numbers also differ between different editions of the same work.

$$A \times B = \{(a, b); a \in A, b \in B\} \tag{3.1}$$

$$M = \{(a, b); a = b, a \in A, b \in B\} \tag{3.2}$$

$$U = \{(a, b); a \neq b, a \in A, b \in B\} \tag{3.3}$$

It is necessary to distinguish between the elements of A and B and the records which describe those objects. In this project, A and B would describe two different collections, probably from two different institutions, with a and b representing the individual physical photographs referred to in those collections. Therefore since the collections originate from different collections, the record creation processes will be different between the two sets. The records of the elements a and b are described as $\alpha(a)$ and $\beta(b)$ respectively. The separate record creation processes mean that if $a = b$ then $\alpha(a) = \beta(b)$ does not hold true in all cases.

Assuming that each element possesses multiple features (e.g. *person*, *date*, etc). It is possible for elements from $\alpha(a)$ and $\beta(b)$ to be identical but not match (i.e. $\alpha(a) = \beta(b)$ but $(a, b) \in U$). The most obvious example of this would occur if all of the features of some elements were empty/absent. Such elements would be identical to each other but should not be placed in set M. Therefore, whilst similarities between $\alpha(a)$ and $\beta(b)$ are an indication of similarity between a and b they do not guarantee it. Conversely dissimilarities may not correspond to dissimilarities between a and b.

There are two possible errors which could occur when trying to compare the individual pairs from $A \times B$. One, a pair (a, b) where $a = b$ is incorrectly placed in set U. Two, a pair where $a \neq b$ is incorrectly placed in set M. There is a, however, third area for concern. When $a = b$ or $a \neq b$ but the record linkage system is unable to determine which set the pair belongs to. Having a large number of unclassified results will reduce the value of the results but not constitute an actual mistake. In most circumstances $|M| \leq |U|$, certainly when comparing GLAM records, the size of the unmatched set will be multiple orders of magnitude larger than the matched one since the number of photographs which are common to multiple collections is expected to be very low.

The expectation that the number of matches available to be found is low is due primarily to the amount of GLAM held material which remains un-digitised. Despite the significant resources applied to the task so far, most heritage artefacts

have not been digitised. A survey of approximately 2000 European institutes in 2012[204] reported that while 88% of institutions have, or are creating digital collections, only ≈22% of their collections have been digitised. The same survey does, however, also report that photographs are the most commonly digitised artefacts, with 32% digitised. The conclusion that must be drawn is that even if the photographs have survived and been passed of a GLAM institute, the odds are very much against them having been digitised and made available online.

The remainder of this chapter examines several potential methods for identifying co-reference and separating record pairs into the matched and unmatched sets. The methods described are the most well established solutions as found from an analysis of the existing literature.

Section 3.1 is devoted to describing expert knowledge techniques (i.e. rule based) with a specific focus on fuzzy logic as a method of handling feature uncertainty.

Sections 3.2 and 3.3 describe supervised learning approaches, discussing feature independence and dependence modelling, as a potential solution to the knowledge engineering problems of rule based solutions.

Finally section 3.4 discusses unsupervised learning approaches in the form of clustering as a potential solution to the training data requirements of supervised approaches. A special focus is placed on fuzzy clustering, which is related back to the Fuzzy Inference System (FIS) systems described as part of the rule based techniques.

3.1 Rule based identification

Rule based co-reference identification is an expert knowledge approach to the problem. A series of rules are created which describe all conceivable circumstances and the action/s to perform in each case. As an expert knowledge approach, it relies on the programmed knowledge of domain experts[238]. Since rule based systems use manually derived rules, they are a heuristic approach[5] and the programmed knowledge can include common sense rules rather than proven mathematical solutions. Therefore while rule based systems may not find the ideal solution for a given set of inputs, they are comparable to human experts and are efficient at

[5]Heuristic approaches use shortcuts and assumptions to produce solutions that are not guaranteed to be optimal but which are 'good enough'. Heuristic approaches are commonly employed when the approach/es for finding the optimal solution have excessive time or resource requirements.

finding good/acceptable solutions.

Rule based systems for commercial applications are popular in part due to their high record throughput and because the internal logic that the systems operate on can be easily examined and modified (in contrast to black box system such as ANNs). This makes the system logic accountable since if pairs are incorrectly classified the cause of the problem can be easily identified. That the internal logic of rule based systems is open in this manner, may prove to be an important consideration when it comes to getting a system accepted by members of the GLAM community.

Since the rules must be manually created, the time required to produce rule based systems is heavily dependent on the complexity of the task. The primary bottleneck in the process is the extraction of knowledge from the domain experts who understand the task and converting this knowledge (which many be difficult to articulate/explain) into the required form. This process is known as knowledge engineering. The result is a series of if then statements (i.e. **IF** x **THEN** y) which represent the rules that a person would use to solve the same task. The resulting sets of x and y for all rules in the rule base are known as the antecedents and consequents respectively[54].

A rule based approach is most successful when the individual pieces of identifying information (features) are identical between co-referent records. For example, phone numbers and postcodes can be reliably represented in a single standard format. Co-reference identification using features (i.e. name) which can be represented in multiple formats or which may have multiple representations introduces a degree of uncertainty in matches between the individual features. Simple rule base co-reference identification is poorly equipped to handle this uncertainty. Feature uncertainty increases the complexity of the expert system since each feature can no longer be modelled as match/no match[149].

3.1.1 Expert systems

Whilst all expert systems are rule based, not all rule based systems are expert systems. A rule or knowledge base is just one of the components that are required[99]:

- User interface - This is just the method by which the user of an expert system

interacts with it. In early expert systems this was a textual[6] interface, but Graphical User Interfaces (GUIs) are now common.

- Knowledge base - The rules known by the expert system, these can either be supplied in advance and, therefore, be a form of background knowledge for the system or they can be supplied by the user during the consultation phase.

- Inference engine - This combines rules from the knowledge base in order to arrive at justifiable outcomes.

- Explainer - Used to describe how the systems output was produced to the user/s, i.e. which rules in the knowledge base were triggered and how they relate to the initial facts supplied by the user.

The major distinguishing feature of expert systems is the presence of an explanation system. Once a decision has been arrived at, an expert system must be able to explain why it arrived at that outcome and which rules and connections it followed to do so[148]. It is generally assumed that these explanations be in the form of natural language explanations and should be understandable without specialised knowledge of how expert systems function[147, 148].

Also important, but not a defining feature of an expert system, is the inference engine. Inference engines attempt to use the knowledge contained in the systems rule base in order to deduce additional facts. Whilst these additional facts can be deduced from the principles stated in the rule base and information supplied by the user, they are not specifically laid out in advance. As such, expert systems do not need every response spelled out, the correct response can be deduced from first principles. Rule deduction can be conducted in either a data or goal driven manner, the difference between the two is whether the inference engine starts with the known conditions and works forwards or it starts with the desired consequents and works backwards in order to see if the conditions do, in fact, support them. These approaches are also known as forwards and backwards chaining respectively.

Expert systems were some of the first attempts at producing true AI and date back as far as the 1960s. The aim was to produce general purpose expert systems able to answer questions regarding almost any domain[13]. These early attempts failed when the sheer size of the rule bases became unmanageable, but by focusing on individual domains or highly related domains the rule bases can be kept sufficiently

[6]Command line.

compact and expert systems still prove effective and remain widely used in real world systems.

DENDRitic ALgorithm (DENDRAL)[65, 125] and MYCIN[31] (which was based on DENDRAL) were the earliest expert systems to achieve notable success and to achieve performances equal to, or greater than, human experts[125]. Importantly, DENDRAL separated the knowledge base from the inference engine and the rest of the system. Treating the knowledge and systems to interpret the knowledge as separate entities, allows each to be expanded and improved independently. Given the lengthy time frame and high resource requirement for knowledge engineering, this potentially allows for the same underlying rule base to be reused in ever more capable systems.

3.1.2 Fuzzy Logic

As mentioned previously in section 3.1, one concern with rule-based systems is that they work best on definite true/false conditions. However, not all real world considerations can be accurately described in this boolean manner[7]. The issues can be mitigated by increasing the number of rules and the rule complexity, but this increases the resources required and the chances for mistakes to be made.

Fuzzy logic is an approximate reasoning technique which uses a multivalued form of logic. Multivalued logics had been under investigation since the early 20^{th} century[169] but fuzzy logic as described here originated with Zadeh in 1965[241]. Fuzzy Logic is one solution to some of the problems found with boolean logic. It allows the outcome of a rule to be true to varying degrees which in turn allows fuzzy logic to handle uncertain and imprecise inputs[141, 242, 243].

What follows is an example of a simple fuzzy rule-based system. In this example fuzzy logic is being used to describe the gratuity to leave at a restaurant[8]. This is an example of a Mamdani FIS, a description of the differences between this and the other established FIS approach (Takagi-Sugeno-Kang (TSK)) can be found in section 3.1.2.2.

The first step is to define the inputs and fuzzy sets, in the example two inputs (*time* waited and food *quality*) are used, each containing three sets. *Fast*, *fine* and

[7]Under a boolean (also called crisp) approach, statements must be either true or false.
[8]Known as the waiter tipping problem, a modified version is used in this thesis[229].

slow for time[9] and *poor*, *normal* and *good* for quality[10]. The time and quality inputs demonstrate that fuzzy logic is able to model not just absolute quantitative values (i.e. time) but also more abstract qualitative concepts (i.e. quality).

When a new query is posed to the FIS the membership value of the input to each fuzzy set is calculated. Using 48 minutes for time under this example would be considered 0.0 fast, 0.4 fine and 0.6 slow[11]. In comparison, if a value of 8 for quality [12] is used then quality would be 0.0 poor, 0.4 normal and 0.6 good.

In order to calculate the output, the input sets are passed through the rule base of the FIS. The output contains a further three fuzzy sets, low, standard and generous[13]. The rules of the example system are:

IF time is slow **OR** quality is poor **THEN** gratuity is low

IF time is fine **AND** quality is normal **THEN** gratuity is standard

IF time is fast **AND** quality is good **THEN** gratuity is generous

Substituting the set membership values into the rule base gives the following:

IF 0.6 **OR** 0.0 **THEN** 0.6

IF 0.4 **AND** 0.4 **THEN** 0.4

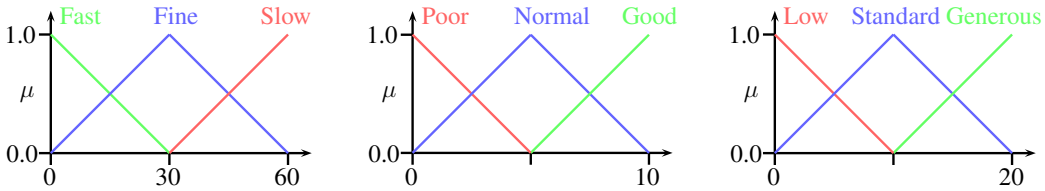
IF 0.0 **AND** 0.6 **THEN** 0.0

Therefore given the input values, the resulting output is 0.6 low, 0.4 standard and 0.0 generous. This is more clearly visualised in figure 3.2.

The final step is the defuzzification of the output sets, this means combining the sets into a single, non-fuzzy value. There are a number of distinct defuzzification methods, and they will be discussed in the next subsection (3.1.2.1). However, the example will be using the centroid approach and the defuzzified output is, therefore, 8.17%.

As the example shows, fuzzy logic allows for rules to be described in a much more natural manner than would otherwise be possible[228]. With traditional non-fuzzy rules, there would either need to be an excessive number of rules in order to accommodate all possible combinations of inputs, or there would be large jumps in

[9] See figure 3.1a.
[10] See figure 3.1b.
[11] See figure 3.2.
[12] A qualitative measure in the range [0 10].
[13] See figure 3.1c.



(a) Time (minutes). (b) Food quality. (c) Gratuity(%).

Fig. 3.1: Input and output sets for and example FIS.

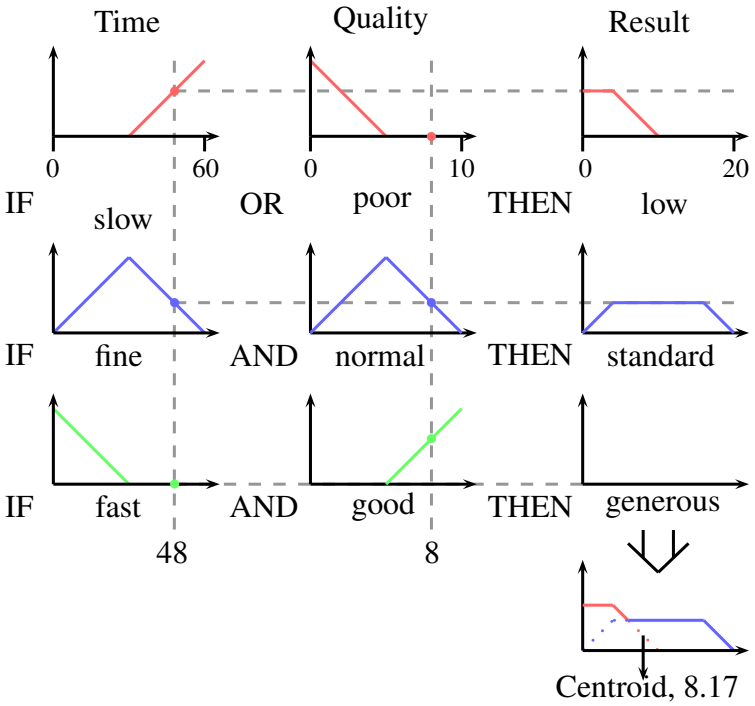


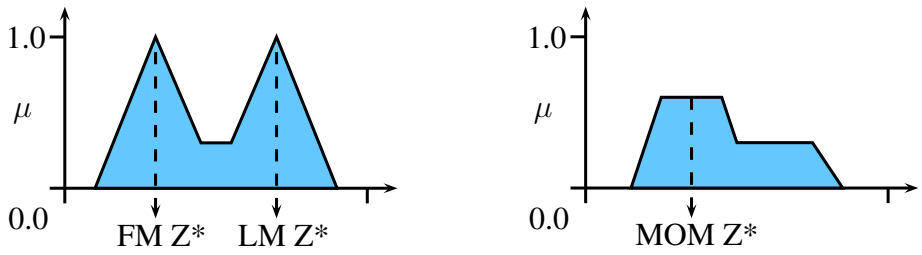
Fig. 3.2: Visualisation of an example FIS.

the output value as one rule after another was triggered. Using a fuzzy system, as the input values change, the output value smoothly changes to reflect even small difference in the inputs[228].

3.1.2.1 Defuzzification

As mentioned in the worked example, the final step is to convert the output fuzzy sets into a non fuzzy value which can be passed to other processes. In the example, this output would be used to control the size of the gratuity. This process is called defuzzification and there are several ways to achieve this. The most common but by no means only methods are[155]:

- Maxima - A simple approach although its application is limited. Takes the x value which corresponds to the highest point in μ. Since the output fuzzy sets may have the same μ value at multiple points along the x-axis, there are several variations which alter which maximum is used. These include first and last maxima, which take the lowest and highest points on the x-axis which achieve the maxima μ value[14], and the mean and median of the maxima. These combine all the points which achieve the μ maxima and then takes the x-axis mean or median of the points. See figure 3.3b where the blue shaded area represents an output set, with each arrow demonstrating the point along the x access which would be returned if the set was defuzzified using the labelled defuzzification approach.



(a) First and last maxima. (b) Mean/median of maxima.

Fig. 3.3: Defuzzification methods for fuzzy sets.

- Centroid - First described by Leszczyski et al.. This defuzzification method finds the centroid or centre of mass of the combined output fuzzy sets[118]. The defuzzified value is simply the position of the centroid along x. See figure

[14]See figure 3.3a.

3.4, this figure follows the same conventions as figure 3.3, however in this figure the centroid point of the set is also included as a circle.

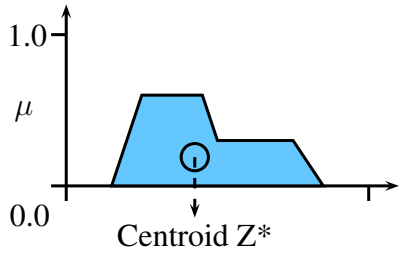


Fig. 3.4: Centroid defuzzification.

- Weighted Average - Which uses the x-axis position of the μ maxima for each of the output fuzzy sets. The x-axis positions are weighted according to the associated μ maxima for each set and the results combined, see figure 3.5 and equation 3.4. See figure 3.5, this figure shows two overlapping consequent sets in blue and teal, the MOM for each set is shown by the arrows and should be combined as shown in equation 3.4.

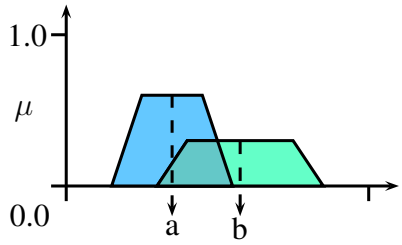


Fig. 3.5: Weighted average defuzzification.

$$weightedaverage = \frac{\sum_{i=1}^{n}(S_{i_\mu} * \mathrm{MOM}(S_i))}{\sum_{i=1}^{n} S_{i_\mu}} \tag{3.4}$$

3.1.2.2 Takagi-Sugeno-Kang (TSK)

The gratuity example described above is an example of a Mamdani FIS. Whilst Mamdani FISs are widely used they are not the only approach to achieve widespread acceptance. Takagi, Sugeno and Kang proposed a different approach[209, 211]. Whilst the initial fuzzification and rules of both approaches are the same, the approaches differ when it comes to the output membership sets and defuzzification. Under the TSK model, the defuzzification of the consequent sets is replaced with a crisp function. This function is typically a polynomial making use of the antecedent set input variables, but potentially any function could be used[101].

These differences mean that Sugeno is more computationally efficient and, therefore, faster than Mamdani, but the rules for Mamdani systems can be described in a more expressive form and, therefore, can more easily model expert knowledge and are more easily understood[108].

The relative quality of the results produced by the two major FIS approaches is still a matter of debate. However the general consensus of the existing literature from a wide range of domains strongly suggests that Sugeno produce the best results if a training data set is available to tune the system[82]. For example, Guney and Sarikaya used a training data set and several optimization algorithms in order to train both Mamdani and Sugeno FISs. The most effective approach they tested was a combination of Sugeno and Least SQuares (LSQ)[53, 138]. The performance of Sugeno systems on average, across all optimization algorithms tested, was also significantly better than that demonstrated by Mamdani.

Deciding between the approaches is, therefore, a case of choosing between performance or comprehension.

3.2 Probabilistic Record Linkage (PRL)

Having examined expert knowledge based techniques, this chapter now turns to reviewing co-reference identification approaches which use supervised learning. This means that the techniques teach themselves how to identify co-referent records using a set of example cases which are supplied in the form of a training data set. Supervised learning does not, therefore, need to be manually programmed with rules and avoids the potentially time consuming knowledge extraction requirements of expert knowledge based approaches.

PRL is a method of comparing records containing multiple fields where matches between those fields have differing levels of confidence. This is achieved by weighting the individual field and allowing each field match to contribute to an overall record match. This approach is often combined with a statistical analysis technique which automatically generates the field weightings. Therefore, whilst PRL is not a supervised learning technique itself, it can be considered one when combined with automatically derived field weights.

First described by Newcombe et al.[153], the approach uses weights attached to the individual elements of the records being compared to identify co-reference

between the records[67]. When comparing two records, the individual elements of the records are compared separately. Should any of the individual elements match, the overall record match increases by the weight value assigned to that element. Finally, the co-referent records are identified by applying a threshold to the record pair match value.

PRL often uses field weights derived by a supervised learning approach run against a training data set. The standard technique is to process the records of the training data in the same way that real records would be. However as a training set is being used, it is already known whether the records being compared actually match. Therefore, it is possible to record how often a match between the individual fields corresponds to an overall match between the records. It is then simple to calculate the probability for an overall record match given a match between a set of elements.

For example, tables 3.1a and 3.1b on the next page show the interactions between a set of four items, a, b, c and d. Each item possesses two attributes denoted in table 3.1a as a_1, a_2 etc. Matches between individual attributes are marked with m while non-matches are marked with u. Overall matches between the items (a, b, c and d) are indicated by green shading. As can be seen in the breakdown of the training data as shown in table 3.1b, attribute 1 matching (m) corresponds to an overall match (M) on 5 occasions with 0 exceptions. A lack of matching by attribute 1 (u) corresponds to a lack of an overall match (U) on 4 occasions with 1 exception. The probability of a match at the level of attribute 1 (m) corresponding to an overall match (M) can therefore be calculated as $M_m/(M_m + U_m) = 5/(5 + 1) = 0.83$. If the same actions are carried out for attribute 2 then the probability is 0.67.

Using these two values, it is possible to identify whether two elements match (M) or do not match (U) just by determining which of their attributes match. For example in the case of a vs. d, as shown in table 3.1a both attributes match (m), therefore the probability values for both attribute 1 (0.83) and attribute 2 (0.67) are summed for an overall value of 1.5 as shown in table 3.1c. However in the case of b vs. d only attribute 2 matches and so the overall value is only 0.67[15]. Table 3.1c on the following page clearly shows that a threshold of > 0.7 and < 0.8 would correctly classify 90% of element pairs. The one incorrectly classified pair is highlighted in red.

[15]Rounded to 0.7 in table 3.1c.

	a_1	a_2	b_1	b_2	c_1	c_2	d_1	d_2
a	m	m	m	u	u	u	m	m
b			m	m	u	u	u	m
c					m	m	u	u
d							m	m

(a) Training data for PRL example.

	M		U	
	m	u	m	u
1	5	0	1	4
2	4	1	2	3

(b) Breakdown of training data shown in table 3.1a.

	a	b	c	d
a	1.5	0.8	0.0	1.5
b		1.5	0.0	0.7
c			1.5	0.0
d				1.5

(c) Co-reference probabilities for PRL example, note that these are non-normalised values.

Table 3.1: Example of training and co-reference identification using PRL.

The central problem with using PRL for co-reference identification is that the technique is effectively a naive Bayes classifier[234] and, therefore, the element matches are conditionally independent from each other[85, 183]. This means that each element is considered completely independent from all others. More complex matches based on the interactions of multiple elements are not possible. Conditional independence means that PRL is unable to detect or model relationships between the individual elements.

Although the ability of PRL to place greater or lesser importances on fields appears to be a promising approach for identifying co-reference, its logic is very simple and may be too simplistic for use with GLAM records. Determining whether or not PRL can be used as part of this research can not be determined in advance and would require practical experimentation.

3.3 Artificial Neural Networks (ANNs)

ANNs are biologically inspired models that contain multiple artificial [16] neurons. Each artificial neuron has a set of weighted inputs, an input function for combining the input signals and an activation function which determines the output strength of the neuron based on the combined input strength. In order to address more complex tasks a neural network may, in the case of MultiLayer Perceptron (MLP) ANNs, have multiple layers[17], each layer contains one or more neurons, taking as their inputs, the outputs of the neurons in the previous layer. Other forms of ANNs do exist[18], but both perceptron and MLP ANNs are feedforward architectures (i.e. information travels in a single direction and the network contains no closed loops) which was one of the first[107] and probably simplest ANN design, other architectures (i.e. RNN) can and do use closed loops.

A significant advantage of ANNs over PRL is that the inputs are not conditionally independent[107]. They are, therefore, able to model more complex problems featuring interactions from multiple inputs. As a secondary benefit, since ANNs learn from a training data set, the expensive and time consuming knowledge engineering requirements of rule based systems are removed.

Depending on the configuration of the systems, ANNs can be classified as both supervised or unsupervised learning systems. In the supervised configuration, a training dataset is used and the neuron weights are modified until the network output matches, or closely approximates, the output shown in the training dataset for a given set of inputs. In the unsupervised configuration, the ANN is expected to tune itself to minimise (or maximise) some performance metric. Use of ANNs as a supervised learning technique is, therefore, dependent on a suitable and available training data set, whilst its use as an unsupervised technique is dependent on a quantifiable output which can be easily (and preferably automatically) scored. Both of these requirements can be problematic. Either a large training dataset is required, which can be a problem when trying to train systems using real world data, or some means of automatically evaluating the output of the system is needed. Since the output in the case of this research would be photographic records, the only viable method of evaluating the quality of the results would be manual

[16] I.e. simulated.

[17] See figure 3.6.

[18] E.g. Radial Basis Function (RBF), Kohonen self-organizing and Recurrent Neural Networks (RNNs).

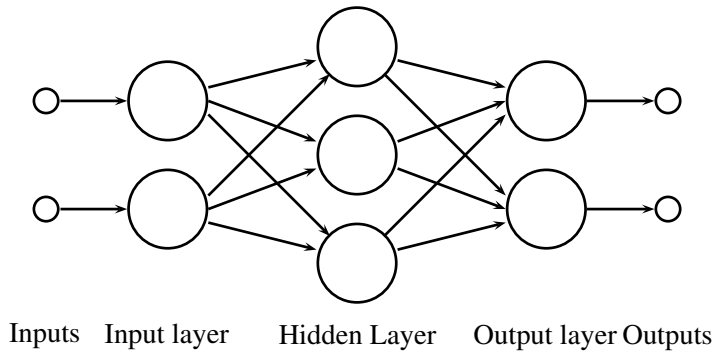


Fig. 3.6: Example of interconnected nodes in a MLP ANN. Note the distinct layers of neurons.

examination by domain experts. Therefore, the use of ANNs as an unsupervised learning technique for this research is very difficult at the present time and the use of ANNs in a supervised capacity remains highly challenging.

A traditional criticism which has been levelled against ANNs, is that they are black box systems. This means that once a ANN has been created and trained, the internal rules of the system are not available for examination. This means that it is not possible to learn why a ANN gives the responses that it does. Whether or not this is a problem as long as the ANN produces the correct results is questionable from a performance standpoint. It does, however, cause problems for transferring learned knowledge from one ANN to another and whether the output will be trusted for real world tasks[112, 216]. For the purposes of this research, some consideration must be given as to how the final approach will be received by the intended users. Whether experts in the GLAM community will accept the results of a search system if it is unable to explain how it arrived at those results is unknown. Whilst keyword based approaches perform poorly, it is at least clear why they returned a specific set of results. Fortunately the extraction of the interval rules from trained ANNs has become a significant focus for ANN researchers. Starting at least ten years ago[216] an increasing number of publications have been made available which present methods of extracting the internal rules from trained ANNs. More recent publications make specific mention of the need to represent the extracted rules in an understandable manner[112].

ANNs offer much in the way of robustness and approximate reasoning and would, therefore, appear ideal for dealing with the imprecise information contained in

GLAM collections. Due to the lack of suitable training data and challenges of automatically evaluating the outputs, ANNs were not considered appropriate for this investigation. However given the advances in knowledge extraction, the traditionally black box nature of ANNs is no longer a major concern and if a suitable training data set could be located/created in the future, ANNs could be favourably reconsidered.

3.4 Clustering

Clustering is an unsupervised learning approach and as such, clustering does not require training data or pre-supplied rules. However if training data is available, such information can be used to tune clustering systems and, therefore, improve performance. Clustering attempts to identify intrinsic structures in the dataset being searched and attempts to separate items into groups of similar items. With regards to co-reference applications, the aim would be to have co-referent records placed in the same cluster and non-referent records placed in different clusters. Since clustering is effective at grouping similar as well as identical items together, it is often used for document classification[19] since it does not require exact matches between the records.

Within clustering methods, there is a wide range of variation between the implementations. However, the vast majority of these can be divided into two main categories; hierarchical and partitional.

3.4.1 Hierarchical clustering

Hierarchical clustering operates by organising the clusterable items into a dendrogram or tree. The most similar items according to whatever distance measure is used to compare the records will be placed in close proximity within the tree. If used for a co-reference system, then the hope would be to see co-referent record pairs ordered as the parent and child nodes of each other. The hierarchical structure produced using this approach means that each cluster is itself built from smaller sub-clusters. This is in direct comparison to partitional clustering[20] where the internal structure within clusters is not easily available. Once the dendrogram has been produced, the hierarchical structure allows for the number of clusters and size of the clusters to be

[19] Grouping together long sections text according to the topic of the text.
[20] See section 3.4.2.

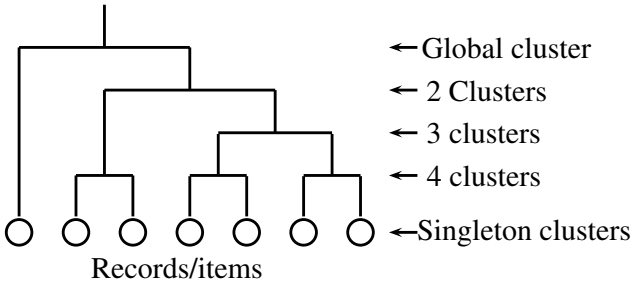


Fig. 3.7: Example of a dendrogram which could be produced by hierarchical clustering.

changed quickly, easily and without recalculation of the dendrogram.

Hierarchical clustering can itself be subdivided into agglomerative or divisive techniques. The distinction between these two approaches consists largely of from which end of the dendrogram the process starts building the tree. With an agglomerative algorithm, the process starts with n singleton clusters (where n in the number of items to be clustered) and merges successive groups of clusters until the desired number of clusters (k) or some other stop condition has been reached. A divisive algorithm starts with k=1 and splits the progressively smaller clusters until singleton clusters are achieved.

3.4.2 Partitional

Partitional clustering is most easily visualised if it is assumed that each record is a point in two dimensional space. The examples presented in this chapter use this approach. Dissimilarity between points is measured as the euclidean distance between them. There is nothing preventing the use of partitional clustering in three or n dimensional space, just as the similarity between the individual records does not have to be the euclidean distance between two points.

One of the common issues with partitional clustering algorithms is that they require a pairwise similarity matrix of the records being clustered. This is not a requirement for all partitional algorithms but when required often causes scaling issues. As the number of records increases, the number of pairwise comparisons required rises exponentially[21]. Therefore, moderate increases in the number of elements being clustered can quickly cause major increases in the processing time and resources required to generate the similarity matrix.

[21]Depending on whether the record similarities are directional or non-directional, the size of a similarity matrix for a set of records (R) will be either $|R|^2$ or $\frac{1}{2}(|R|^2 + |R|)$.

Many partitional algorithms require that the number of clusters to be found is specified in advance. This means that changing cluster numbers or cluster sizes is computationally expensive when compared to hierarchical clustering.

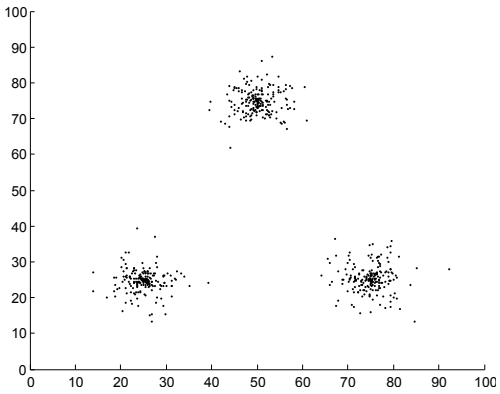
3.4.3 Description of k-means

One of the most common clustering approaches is k-means[132]. The standard implementation of which uses an iterative refinement approach to identify a pre-specified number of clusters[127].

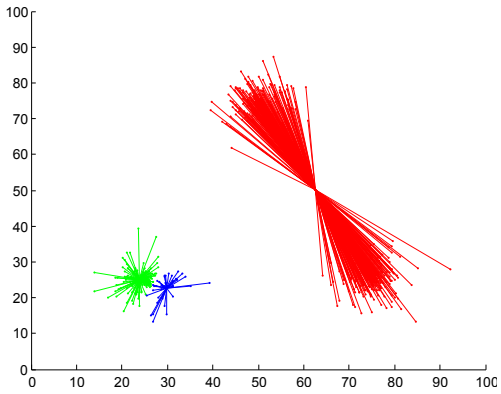
Given an initial distribution of k cluster centroids, assign each point in the search space to the nearest one. Nearest is defined as the centroid with the lowest Euclidean distance to that point. The centroids are now repositioned as the mean of all points in that cluster. New centroid positions continue to be generated for either a present number of iterations or until the change in centroid positions between iterations falls below a pre-set threshold.

Whilst this approach is simple to implement, it does have several issues. These are not limited to just k-means clustering. As one of the simplest clustering algorithms, k-means is susceptible to the widest range of issues. The major problems are:

1. The clusters produced are sensitive to the initial cluster centroid distribution. As can be seen in figure 3.8, poor initial placement has caused one cluster to be identified as two separate clusters (green and blue), whilst the red cluster has combined two distinct groups of points into one.



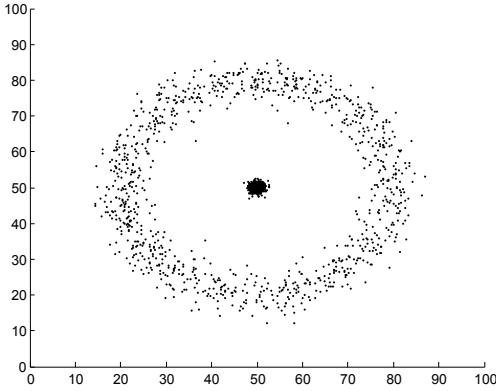
(a) Distribution of points in three equally sized and spaced clusters.



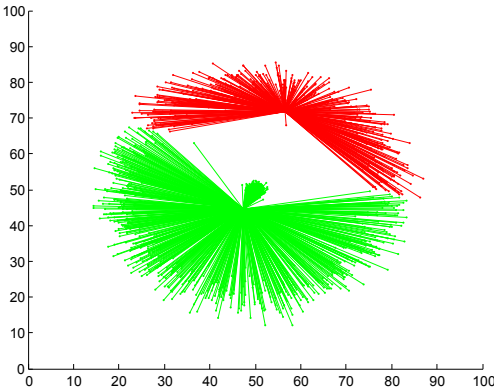
(b) k-means identified clusters, $k = 3$.

Fig. 3.8: Example of poor initial centroid placement for k-means clustering.

2. Non-globular clusters will not be correctly identified using k-means or its variations. An example of this is shown in figure 3.9. As can be seen in figure 3.9, one small cluster is ringed by a second larger cluster. Instead of treating clusters separately, k-means has split the larger cluster in two and combined one of these outer ring clusters with the central points.



(a) Distribution of points in one globular and one non-globular cluster.



(b) k-means identified clusters, $k = 2$.

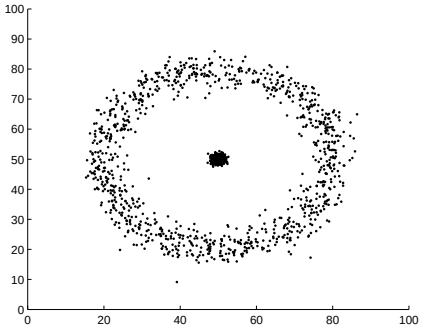
Fig. 3.9: Example k-means clustering on concave clusters.

3. Since the first step is to generate k centroids this means that the value for k must be known in advance. If the number of clusters expected or desired from the algorithm is not known, an additional process is required in order to generate a k value (e.g. gap statistics[215], X-means[168] or VAT/iVAT images[22]). The point distribution resembles that of the non-globular example. Whilst in this case, the central points are correctly identified as being a separate cluster, the outer ring is now incorrectly split into four distinct clusters.

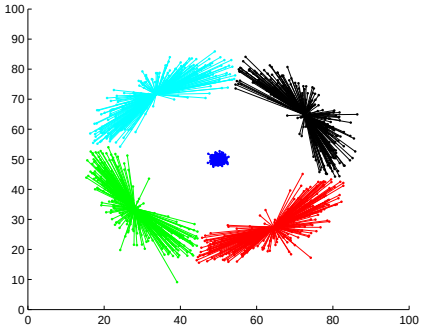
4. k-means makes no allowance for clusters of differing sizes or densities. An example of the classification problems that this can cause is shown in figure 3.11. In the example, the approximate positions of the three clusters are correct. However, points which belong to the larger, diffuse cluster are being incorrectly combined with the smaller, denser clusters.

5. Calculating the centroids is computationally expensive for large datasets[168]. As the number of items to be clustered increases, the number of distances to be calculated increases quadratically.

[22] See section 3.4.4.1.

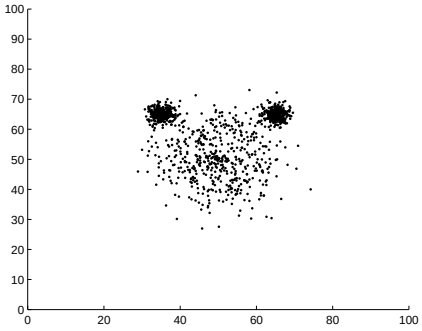


(a) Distribution of points in one globular and one non-globular cluster.

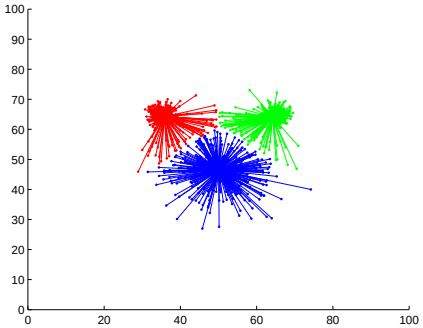


(b) k-means identified clusters, $k = 5$.

Fig. 3.10: Example of an unsuitable k value for k-means clustering.



(a) Distribution of points in two small dense clusters and one large diffuse cluster.



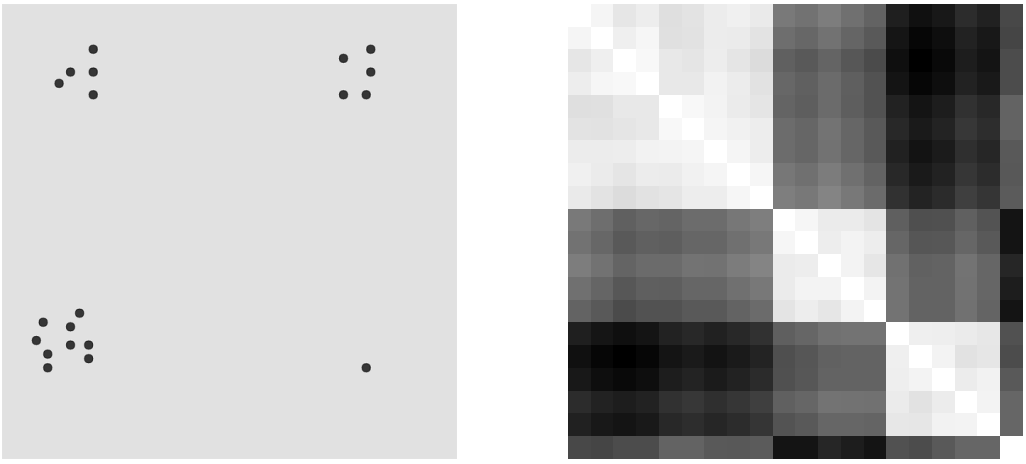
(b) k-means identified clusters, $k = 3$.

Fig. 3.11: Example of k-means clustering on clusters of difference sizes/densities.

3.4.4 k requirement

As previously stated, some clustering algorithms require that the number of clusters be provided as an initial variable to the algorithm. This can cause problems if the number of clusters in the data is not known. One solution would be to use a density based algorithm instead and, therefore, remove the need for a k value[23]. If that is not an option, there exist data analysis tools which can be used identify promising k values from a data set.

[23]Or C value in the case of Fuzzy C-means, see section 3.4.6.1 on page 56.



(a) Data points. (b) VAT image.

Fig. 3.12: Example data points and corresponding VAT image.

3.4.4.1 Visual Assessment of cluster Tendency (VAT) images

One method of determining suitable k values is to use VAT images[25]. VAT images are square images containing n^2 pixels where n is the number of items/records and each pixel represents the dissimilarity value of a single pair of records; as such it is simple a visual representation of a dissimilarity matrix. With the pixels places in random locations, VAT images would convey little useful information. Therefore, the VAT algorithm reorders the records in the dissimilarity matrix so as to place the records with the lowest dissimilarities adjacent to one another. The result is an image like the one shown in figure 3.12b; a series of black squares of varying sizes indicate not just the number of clusters in our example data[24] but also the likely sizes of said clusters.

k estimation is not the only possible use for VAT images. The images produced also offer a useful visual analysis of the suitability of a data set for clustering. They can be used to investigate if there is sufficient intrinsic structure in a data set for clustering to be worth attempting (more often referred to as assessing cluster tendency). This is important since algorithms such as k-means will produce clusters regardless of whether there are, in fact, suitable groups in the data. VAT images are not a guarantee of clustering suitability, but they are a promising indication as well as being significantly simpler to produce computationally than actual clusters.

[24]See figure 3.12a.

3.4.5 Density based clustering

Density based clustering algorithms attempt to group nearby points as with all clustering algorithms. However, density based clustering algorithms take into account the relative cluster densities to which points are assigned. This means that density based approaches are better able to identify clusters of differing sizes, especially when the clusters are in close proximity (see figure3.11). Examples of this approach include Density Based Spatial Clustering of Applications with Noise (DBSCAN), CLustering In QUEst (CLIQUE), Merging of Adaptive Finite IntervAls (MAFIA), Fuzzy clustering by Local Approximation of MEmberships (FLAME) and DENsity CLUstEring (DENCLUE)[10].

3.4.6 Hard vs. fuzzy

The membership of records to clusters can be described in two different ways. The simpler approach is to force a hard record membership (i.e. k-means). This means that records belong fully to a single cluster. The second approach is a fuzzy membership (i.e. Fuzzy C-Means[24, 58]). Following on from the concept of partial set membership described in section 3.1.2 on page 40, under a fuzzy approach records can belong to one or more clusters to varying degrees of membership. An alternative but equally valid view is that all items are members of all clusters although that degree of membership can be zero. The origin and concept of applying fuzzy logic to clustering can be seen in the work Ruspini and Bellman et al.[22, 75, 188].

Fuzzy memberships can be converted to hard memberships simply by assigning each item to the cluster with which it has the greater membership, but conversion in the other direction is not possible.

3.4.6.1 Description of Fuzzy C-means

This approach is very similar to k-means both from a conceptual and implementation standpoint. Since every point can belong to every cluster, the centroid calculation must be updated or else they will all converge on the same location. The updated calculation takes into account the likelihood of each point belongs to the cluster surrounding that centroid and uses that likelihood to weight that point's contribution. This likelihood of belonging is inversely related to the distance of the point to the centroid's current position.

3.4.7 Post clustering processing

As with many processes, the results produced by clustering can be improved or at least cleaned up with post processing of the output. Common examples include the removal of small/singleton clusters as these can often be attributed to noise in the input data[90, 94, 105].

3.5 Conclusions

All of the approaches described above have their own unique advantages and disadvantages. For this project however, the major determining factor must be the training data requirements. Whilst suitable training sets do exist for co-reference identification in other domains, in particular in the area of document classification. These are not suitable for the GLAM records that this project is looking at or GLAM catalogues more generally. This is because the records in the available document classification training sets contain significantly more text than is available from GLAM records, both in ERPS or in general. For example, document classification tasks are already served by the following data sets; 20_newsgroup (messages from twenty Usenet message boards), NLM_500 (PubMed documents), Reuters Corpora (Reuters Ltd news articles)[121, 160], TREC datasets (multiple datasets covering a range of topics)[161], and Wikipedia[70, 74]. At the present time, no pre-existing data set which resembles GLAM records and could, therefore, be used for training is apparent. This research is, therefore, conducted under the assumption that none exist.

Creating a new training data set is not an option since in order to create one using real world data it would be necessary to perform a manual co-reference search on thousands of records. As previously stated in section 1.1, the time and resource requirements of a manual search of the ≈34,000 ERPS records is considered unacceptable. In a questionnaire of eight photo-history researchers[25] an average search time of ten minutes per search was reported[26]. Accepting that this value was pro-

[25] See section 7.1.1 on page 132.

[26] Average search time was calculated by dividing the total amount of time spent searching (380 minutes) by the number of records searched for (38). Seven individuals searched for 5 records, one searched for 3. Maximum time spent searching by an individual was 80 minutes (for five records), the minimum was 15 minutes (for five records).

duced from a small, self reporting sample[27], that search times in humanities domains other that photo-history may differ and that during the testing the size of the search space was constrained[28]; the ten minute value provides a starting point from which to work. If searching for all of the missing images was actually attempted the process would likely take longer, in part due to the large search space and in part due to researchers not being able to maintain a 10 minute per search pace continuously.

If all of the ERPS images with 'missing' images were searched for, it would take one person more than two and a half years[29] to conduct the 33,157 searches. Despite the significant achievement those searches would represent, certain neural networks applications can require 400 million training records[143]. While the task of co-reference identification in GLAM collections may not need a training data set of that magnitude, the example does help to highlight the very significant and unrealistic time and resource requirements that would be required in order to create a suitable training dataset from scratch.

Creating a training set using artificial records based on the formats and patterns seen in real GLAM records is a possibility. However, ensuring that the training data set was an accurate representation of the real records would be very challenging and would still be time consuming.
Given the difficulty in obtaining a suitable training data set it is concluded that supervised learning approaches, whilst likely effective, are not suitable for comparing the overall records. They may still prove useful in comparing the individual fields of the records, but this seems unlikely given the time and resources which would be needed to create suitable training sets, even for individual fields. Therefore, this research will only be exploring unsupervised and expert learning systems.

The knowledge extraction aspect of rule based systems if an expert knowledge approach is utilised is still a concern for two reasons. Firstly, whether the mental rules used by GLAM experts can be described and lend themselves to encoding in a knowledge base is unknown. Whilst various methods have been developed to assist in this transfer of knowledge[88, 99] this remains a potentially lengthy

[27]Self-reported responses can suffer from both over and under reporting depending on circumstances and have therefore issues regarding accuracy.
[28]The researchers were limited to searching just five collections.
[29]2.7 years. Assuming one search every 10 minutes, an 8 hour work day and no days off.

process[30][99]. Secondly, due to the somewhat interpretive nature of photographic study, there may be disagreements between different experts as to which elements and combinations of elements constitute evidence of co-reference. This would complicate the development of an acceptable rule base. In order to avoid these issues, an unsupervised learning approach looks most promising. Specifically a partitional clustering approach where the similarities between each field can be considered as a distance in an n-dimensional space appears a promising and logical way forwards. However elements of other approaches such as PRL may be of use in, for example, weighting the importances of the various fields given that fields such as *title* are expected to play a greater role than others, e.g. *date*[31].

This chapter has shown that co-reference identification[32] is a long established problem with multiple techniques and potential solutions already existing across a variety of domains. The approaches which are reliant on training against a pre-existing training datasets are, however, unlikely to be employed due to the lack of said datasets and lack of resources available to create one. This research will, therefore, need to focus on approaches that do not use machine learning.

[30]Knowledge extraction has been referred to as the "bottleneck problem" for expert systems, in part due to the time required for it[64].

[31]Based on the responses to the investigatory questionnaire, see section 1.2.

[32]Or record linkage, entity resolution etc.

4

Text comparison

Whilst the previous chapter described various methods for identifying co-reference between records, all the processes described relied on it being possible to compare the individual fields in the records being examined. Given the nature and contents of the GLAM records, comparing fields in this research project mainly means being able to compare text.

This chapter discusses methods of approximate text comparison. This means both the comparison of individual terms but also of whole sections of text. Approximate comparison in the first case mainly refers to comparison in the face of spelling mistakes and variations. Approximate in the latter case refers to comparison of the meaning of the text. So that texts which describe the same subject can be identified, even if different terms are used to do so. Both forms of comparison will have a role to play in comparing the GLAM record fields.

4.1 Approximate string comparison

In an ideal world, information held by GLAM institutions could be assumed to have been spelt correctly. However, given the quantity of information held, mistakes are a near certainty. Exact string[1] matching (comparing strings on a character by character basis) is the standard method of comparing words in text. Given errors in the GLAM records however, exact string matching would miss valid matches in the record text when one of the records contained a mistake in the text.

Mistakes in the records of GLAM collections can be traced back to three main causes. These are:

[1]Within computer science "strings" refer to a data type containing a series of characters and are how text is stored and manipulated in software.

1. Typographical errors[2] - Mistakes made when creating new digital records or transcribing physical records into digital surrogates (E.g. erps33398[3], "Deisgn". These can be a result of typing mistakes or the physical record being unclear. Various methodologies do exist which can prevent or reduce the number of these errors which make it into the published records. However, these reduce the rate at which records can be digitised as well as increasing the cost and are, therefore, typically only used in safety critical situations such as medical data.

2. Policy - Depending on the policies of the individual institutions, mistakes in the digitised records may be an accurate representation of mistakes in the physical record which was transcribed. For example, the ERPS collection records attempted to record accurately the information held in the original exhibition catalogues, mistakes included. Additional mistakes may have crept in during the digitisation and metadata creation processes, but many of the typographical errors found in the records are also present in the physical catalogues. For example erps29393[4], in which the word "Sunshine" is recorded as "Sunsh¨ne" and in erps12574[5] the word "Atmosphere" appears as "Atmospherie" in both the physical and digital records.

3. Conversion - Given the quantity of information to be digitised it is unsurprising that the relevant digitisation projects can and have taken years to achieve. In that time the software in which the information is stored has been upgraded, replaced and the information has been copied, transferred, migrated and updated. The passage of the records through a variety of formats, software packages and automated processes can leave traces in the data. For example, the ERPS records contain occasional HyperText Markup Language (HTML) tags which were not part of the original physical records. Other collections show similar HTML, mangled Unicode characters and evidence of Comma Separated Values (CSV) formatting.

These issues apply to all record fields although the impact that such mistakes may have varies. For example, errors in the *date* field can be more easily spotted

[2]Typos.
[3]`http://erps.dmu.ac.uk/exhibit_details.php?etid=100605`
[4]`http://erps.dmu.ac.uk/exhibit_details.php?etid=100069`
[5]`http://erps.dmu.ac.uk/exhibit_details.php?etid=132854`

whilst errors in the *person* field are particularly difficult to identify and correct given the level of possible variation in the spelling of names[6].

Given the known existence of mistakes in the record fields, field comparison could prove difficult. This section reviews various approximate string comparison metrics which can be used to compare words despite potential typographical errors. As an added advantage, these metrics can also be effective at word comparison despite regional spelling differences (i.e. 'colour' vs. 'color').

Phonetic approaches are discussed first before moving on to edit distance based techniques. Finally examined are miscellaneous techniques that resemble edit distance based approach but which do not (solely) use edit distance.

4.1.1 Phonetic

Phonetic similarity metrics attempt to compare words according to the similarity of their spoken forms[245]. In practice, this means that phonetic algorithms are most effective when analysing words from a single language and often region. In order to achieve the best results, phonetic algorithms need to be tuned in order to model the specific accents and pronunciation of individual regions[150]. This can limit the usefulness of phonetic approaches.

4.1.1.1 Soundex

Widely considered to be the original approximate string comparison metric, Soundex originated in 1918[156] and continues to be widely used/implemented today. Its usefulness as a generalised string comparison metric today can be limited, as a comparison using this method produces only True or False matches with no intermediate values. It is not, therefore, possible to tune the sensitivity of the approach to varying levels of record quality.

The Soundex algorithm converts strings into a Soundex code. The rules of the algorithm ensure that homonyms and similar sounding words will produce the same codes. Soundex codes are four characters long and consist of a single letter followed by three digits (e.g. A123). For example, under American Soundex 'raven', 'riven' and 'ripen' are all encoded as R150. Originally designed to compare surnames and organise records into general groups with the final name comparison being done

[6]E.g. 'Shawn', 'Sean' and 'Shaun' or 'Steven' and 'Stephen'.

Value	Characters
1	b, f, p, v
2	c, g, j, k, q, s, x, z
3	d, t
4	l
5	m, n
6	r

Table 4.1: Character values for the Soundex algorithm.

manually, as such Soundex's ability to distinguish between long words with distinct endings is poor.

The original Soundex algorithm[7] is as follows:

1. If a pair of adjacent characters would have the same value according to table 4.1, remove the second character.

2. Starting from the second character in the string, remove all occurrences of vowels, 'y', 'h', and 'w'.

3. Starting from the second character, replace the characters with the values shown in table 4.1.

4. The Soundex code is the first four characters of the resulting word. If the resulting word is less than four characters then pad the result with "0"s until it is.

American Soundex is perhaps the most widely implemented and is a good exemplar. However, other versions do exist and will often change the character encoding values shown in table 4.1. For example, the Daitch-Mokotoff rules handle 69 possible characters or character combinations with the intent of modelling the phonetics of Slavic and Yiddish surnames.

4.1.1.2 Alternatives

Soundex was merely the first of a number of phonetic algorithms. Several alternatives of increasing sophistication and complexity have been created since which have produced ever more complete phonetic modelling. Metaphone, for example,

[7]Often called American Soundex.

is the second most widely established phonetic approach[8], but there are many others including Henry Codes (Suitable for French pronunciations), Caverphone (New Zealand pronunciations)[96], NYSIIS[210] (New York State Identification and Intelligence System, American pronunciations), Klner Phonetik[236] (German) and Nominex[199] (British).

What is clear from the multitude of different techniques and rule sets for Soundex is that, if a phonetic algorithm is used, it must be carefully selected for the specific language used in the text and region in which it originated. Given the international nature of GLAM collection records, no matter which phonetic algorithm is selected, it will perform poorly on some of the records. The use of phonetic algorithms as part of this research is, therefore, unlikely.

4.1.2 Edit distance

In comparison to phonetic approaches, edit distance based techniques model string similarity in terms of the number of changes (edits) that are required in order to convert one string into another. Edits include:

- Insertion - inclusion of an additional character within the string (e.g. 'bat' → 'boat').
- Deletion - removal of a character from a string (e.g. 'boat' → 'bat').
- Substitution - replacement of one character with another (e.g. 'bat' → 'cat').
- Transposition - swapping the positions of two adjacent characters (e.g. 'freind' → 'friend'). Whilst transposition errors are one of the most frequent forms of errors made, transposition is the rarest form of edit when it comes to algorithm inclusion. This is because it can also be achieved using a combination of insertion and deletion, though this requires more edits and effects the similarity values generated.

Which of these edits steps are used depends on the specific matching algorithm.

These approaches are highly effective at identifying similarities between mistyped words since the vast majority of spelling mistakes are a matter of accidental in-

[8]Although this covers several distinct versions in the form of Metaphone, double Metaphone (Metaphone 2)[172] and the commercial Metaphone 3[171].

sertions, deletions, substitutions or transpositions rather than homonym based errors[9][61].

4.1.2.1 Hamming distance

The Hamming distance algorithm compares two strings of equal lengths. By counting the number of positions in the strings where the characters are different, the Hamming distance describes how many substitution edits would be required to convert one string into another[84]. For example, $hamming(\text{'cat'}, \text{'bat'}) = 1$ and $hamming(\text{'birch'}, \text{'bench'}) = 2$.

Hamming distance is a measure of dissimilarity, the value increases as more differences are identified. Given that this process can only compare strings of equal size it is of limited use for general text comparison and is more commonly used for error estimation although it is also effective at identifying substitution and transposition errors (hamming distances of 1 and 2 respectively). Whilst it is very unlikely to be used as part of this research, it is included here as it is the simplest possible edit distance based similarity metric and provides a simple example of edit distance.

4.1.2.2 Levenshtein and Damerau-Levenshtein distance

Levenshtein distance was originally proposed and named after Levenshtein in 1966[120]. It allows for addition, subtraction and substitution operations. The approach was later expanded by Damerau to become Damerau-Levenshtein with the inclusion of transposition operators[49]. The effect that this additional operator

[9] Analysis based on GCSE English work, GLAM collection records are expected to show some differences in the types of the mistakes that appear but a specific analysis of GLAM records is not available.

has on the edit distances of various typographical errors can be seen below[10]:

$$\begin{aligned} \mathrm{L}(cat, cats) &= 1, cat \rightarrow cats \\ \mathrm{DL}(cat, cats) &= 1, cat \rightarrow cats \end{aligned}$$

$$\begin{aligned} \mathrm{L}(colour, color) &= 1, colour \rightarrow color \\ \mathrm{DL}(colour, color) &= 1, colour \rightarrow color \end{aligned}$$

$$\begin{aligned} \mathrm{L}(test, testing) &= 3, test \rightarrow testi \rightarrow testin \rightarrow testing \\ \mathrm{DL}(test, testing) &= 3, test \rightarrow testi \rightarrow testin \rightarrow testing \end{aligned}$$

$$\begin{aligned} \mathrm{L}(tpyo, typo) &= 2, tpyo \rightarrow tpo \rightarrow typo \\ \mathrm{DL}(tpyo, typo) &= 1, tpyo \rightarrow typo \end{aligned}$$

The effect of allowing transposition edits on the dissimilarity values produced is shown by the last example where $\mathrm{LD}() = 1, \mathrm{L}() = 2$. A simple, but relatively inefficient recursive example of the Levenshtein algorithm is shown in algorithm 1.

Unlike Hamming distance, Levenshtein and Damerau-Levenshtein distances demonstrate an effective and popular method for comparing terms despite the existence of typographical errors or regional spelling variations[11].

4.1.3 Edit distance resembling approaches

Some term similarity measures use neither phonetic or edit distances. Whilst these algorithms do not use edit distances to calculate string similarity directly, they are generally close relations to those which do (for example, Jaro indirectly uses the number of transpositions). They are, therefore, sometimes called edit distance resembling approaches in the literature[41].

4.1.3.1 Jaro and Jaro-Winkler

Developed by Jaro[102], it was further expanded by Winkler in 1990. As such, the two approaches are mostly identical. Although they do not calculate the number of edits which would be required in order to convert one string into another, by

[10] Levenshtein $L(x, y) = dist$, Damerau-Levenshtein $DL(x, y) = dist$

[11] See section 2.1 on page 19.

Algorithm 1 Levenshtein distance.

Input: Pair of strings to be compared, S and T.
Onput: Number of edit operations to convert S into T.

```
procedure LEV(S,T)
    if S_0 != T_0 then
        cost ← 1
    else
        cost ← 0
    end if

    if |S| = 0 then
        return |T|
    else if |T| = 0 then
        return |S|
    else
        return min( LEV( S_{1,...,|S|}, T ) + 1,          ▷ deletion edit
                    LEV( S, T_{1,...,|T|} ) + 1,          ▷ insertion edit
                    LEV( S_{1,...,|S|}, T_{1,...,|T|} ) + cost )   ▷ substitution edit
    end if
end procedure
```

counting the number of likely transpositions they do factor in one element of edit distance. The change made by Winkler places a preferential weighting on the start of the strings to be compared. This means that the start of the strings is more important than the end of the strings, i.e. 'cat' and 'cab' have a higher similarity than 'cat' and 'bat' despite both being only one character different.

$$\text{jaro}(s_1, s_2) = \frac{1}{3} \cdot \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) \tag{4.1}$$

$$n = \frac{max(|s_1|, |s_2|)}{2} - 1 \tag{4.2}$$

$$\text{jarow}(s_1, s_2) = \text{jaro}(s_1, s_2) + (max(l, 4) \cdot p \cdot (1 - \text{jaro}(s_1, s_2))) \tag{4.3}$$

The equation for the Jaro distance is shown in 4.1. Jaro accepts two strings s_1 and s_2 as inputs. m is the number of matching characters between s_1 and s_2; matching characters are defined as those characters which have a matching character in the other string and are separated by no more than n characters. The value of n varies with the lengths of s_1 and s_2, the relevant calculation is shown in equation 4.2. t is the number of transpositions in the matching characters and is,

therefore, affected by the value of n.

Equation 4.3 shows the addition to the original Jaro algorithm which was made by Winkler. l in this case is the number of matching characters at the start of the compared strings. l is normally limited to $l <= 4$. p is the scaling factor which controls the importance which is placed upon the first l characters, Winkler used a constant of $p = 0.1$[237]. p should always remain $<= 0.25$ or else it is possible for the final result to be > 1.0.

This beginning of the string preference means that Jaro-Winkler is particularly well suited for comparing names (both person and place) since shortened versions (e.g. 'Matthew' vs. 'Matt', 'David' vs. 'Dave') will typically only change the later characters (although exceptions do exist, e.g. 'Richard' vs. 'Dick'). When used for comparing person names, Jaro-Winkler has been demonstrated to achieve a real world performance of 97.4% and 97.7% (negative and positive link accuracies respectively)[78]. Jaro and Jaro-Winkler are, therefore, effective methods of term comparison in cases of typographical errors or regional spelling variations and are particularly effective for name comparison where shortened names are a concern.

4.1.4 Conclusions

Approximate string comparison algorithms offer a valuable means for comparing field information despite the errors which are sure to exist. While not all GLAM record fields contain textual information (e.g. *date*) the majority do, and so approximate term comparison techniques are certain to be deployed as part of this research.

The majority of these approximate comparison methods return a similarity value, as opposed to a boolean answer, therefore static thresholds are often employed in order to determine if strings are sufficiently similar to constitute a match. Arriving at these threshold values will be a matter of experimentation and intuition since suitable training data is unavailable.

Whilst the general usefulness of phonetic algorithms for certain approximate matching tasks has been established[170]. Given the entirely written nature of the records to be analysed, a need for them within this project cannot currently be foreseen. There are two main arguments behind this conclusion.

Firstly, the continued development of phonetic algorithms[171] and the sheer number of variations on the original algorithms, each of which is tuned to a specific

language or region, clearly demonstrates that the modelling rules used by phonetic algorithms are currently unable to model the real world complexity of multiple languages/accents. This is in marked contrast to the stable nature of most edit distance and edit distance like algorithms. Whilst the majority of the ERPS collection information is in English, it does contain some non-English words (E.g. erps33372[12], "Coup de Soleil" or 'sunburn' in English)[13]. Whichever variations of the fundamental phonetic algorithm is considered, they will, therefore, perform poorly on at least some of the text.

Secondly and more importantly, phonetic approaches are best suited to situations where the strings contains homophone[14] errors. They are not suitable for identifying the most common spelling mistakes[15][61]. For example, Soundex's use of the first character of the string as the first character of the resulting Soundex code means that it is completely useless if the first letter is incorrect (e.g. 'smith' and 'msith')[166]. Edit based distance approaches have the definitive advantage when it comes to comparing strings containing spelling mistakes, as well as performing well in cases of homonyms and homographs. This is not specific to photo-history and/or GLAM records and so the continued use of techniques such as Soundex in other systems is likely due to it's ease of use/implementation[166] and potentially a lack of knowledge about more capable alternatives.

Therefore, whilst phonetic algorithms are considered too limited for this research, edit distance or edit distance resembling algorithms are seen as being vital for this research. Jaro-Winkler appears to be the logical choice for comparisons in the *person* field given its established performance in name comparison, but Levenshtein or Damerau-Levenshtein or could still play a role in comparing the *title*, *description* and/or *process* fields.

4.2 Textual similarity

Whilst the approximate string metrics discussed in section 4.1 are useful for comparing individual terms, in order to compare fields such as *title* and *description* it

[12] `http://erps.dmu.ac.uk/exhibit_details.php?etid=100579`

[13] It is not easy to determine the exact number of non-English word in the ERPS collection. Attempts to identify non-English words using dictionary lists failed due to the significant number of names and technical photography terms. However, an analysis of 200 randomly selected records found that just 2 (1%) of records contained non-English words if non-English, but correct, place and person names were not counted.

[14] Same sound but different spellings, e.g. 'sea' vs. 'see', 'there' vs. 'their' vs. 'they're'

[15] I.e. transpositions.

will be necessary to compare whole sentences.

Many of the methods described below are, in fact, general vector comparison metrics and can, therefore, be used to compare feature vectors containing any form of numerical data. For the purposes of this section however, textual examples will be used to demonstrate the various methods and discussing their potential advantages and disadvantages.

4.2.1 Term Frequency (TF) and TF-Inverse Document Frequency (IDF)

In order to use a vectorial similarity approach to compare pieces of text, those pieces must first be converted into a vectorial form. In the cases of cosine or Okapi BM25, word order within the text is not taken into consideration. The only features which need to be represented in the vectors are which terms appear in the text and the number of times that they appear. A vectorial representation of this text in this format is widely referred to as a term vector. The number of times that a word appears in a piece of text is known as the Term Frequency (TF). Although the TF values can be used in term vectors directly, the value is often modified in order to take into account the term's usefulness in classifying similar texts. Certain words are so common as to remove any use that they might have as part of a keyword search, good examples of these are articles (e.g. 'the', 'a', 'some') and prepositions (e.g. 'to') although there are others[16]. These terms provide very little (if any) useful information in the context of a search system as proven by the automatic removal of many of these terms when they are supplied to search engines[17]. IDF is the most commonly described approach. $\mathrm{Idf}(t, D)$ is calculated as the log of the total number of documents in the search space D divided by the number of those documents containing the term t (see (4.4)). Without the IDF value weighting term importance the sheer number of appearances of these terms overpower keywords of actual significance.

TF and Term Frequency-Inverse Document Frequency (TF-IDF) only measure the number of occurrences of a single word in a piece of text, in order to measure the similarity of two pieces of text, the term vectors as a whole need to be compared.

[16]In the case of the ERPS records 'photograph' would be one, appearing in 746 (72%) of the 1,040 ERPS records with visual representations.

[17]E.g. Google.

$$\text{idf}(t, D) = \log \frac{|D|}{|D_t|} \tag{4.4}$$

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D) \tag{4.5}$$

4.2.2 Binary vector methods

There are several methods which treat text as a binary vector, this is to say that instead of measuring the number of times that an individual word appears in the text, the appearance or non-appearance of a word is recorded as either true or false. A pair of binary vectors can then be easily compared using methods such as[136]:

- Matching Coefficient - the number of terms which appear in both vectors.

$$|A \cap B| \tag{4.6}$$

- Dice/Sørenson Coefficient

$$\frac{2|A \cap B|}{|A| + |B|} \tag{4.7}$$

- Overlap Coefficient - the same as the matching coefficient but expressed as a proportion of the length of the smallest of the vectors.

$$\frac{|A \cap B|}{min(|A|, |B|)} \tag{4.8}$$

- Jaccard index

$$\frac{|A \cap B|}{|A \cup B|} \tag{4.9}$$

All of these methods could also be used for approximate string comparison of individual terms[18], and some (i.e. Jaccard index) are mentioned for this purpose in the literature. The performance of these approaches is, however, much lower than of more sophisticated measures such as Jaro as they give no consideration to the ordering of the letters or number of occurrences of each letter within the words. Therefore, they are unlikely to be used for term comparison as part of this research.

[18] As previously discussed in section 4.1.3.

4.2.3 Cosine similarity

Cosine similarity is the most common measures of the similarity between two text vectors, in no small part due to its computational efficiency and intuitive nature compared to other approaches[136]. Similarity in this case is defined as the cosine of the angle between the vectors[137].

$$sim = cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{4.10}$$

$$\|A\| = \sqrt{{A_1}^2 + {A_2}^2 + \cdots + {A_n}^2} \tag{4.11}$$

$$cos(A, B) = \frac{A \cdot B}{\sqrt{A^2 \cdot B^2}} \tag{4.12}$$

Cosine similarity is easiest to visualise the example is restricted to two dimensions (and consequently to vectors containing only two elements). However, cosine similarity can be used to compare vectors of any size, each additional element (or word) means that the vector exists in one more dimension. Figure 4.1 shows the vectors described in table 4.2 plotted as lines on a graph. The attributes of the various vectors are simply the TF of each term within the text. For example, the text corresponding to vector A contained three occurrences of one term and nine occurrences of the other.

	Attributes	
	n_1	n_2
A	3	9
B	5	5
C	9	3

Table 4.2: Term vectors for cosine example.

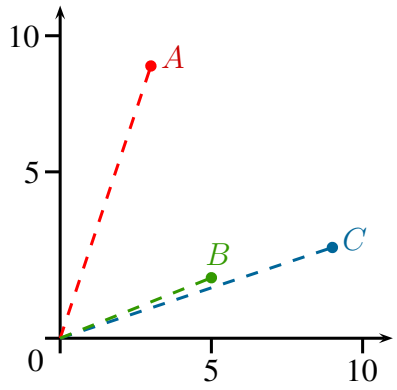


Fig. 4.1: Plotted term vectors from table 4.2.

Since similarity is measured as the angle between the vectors, the differences in the magnitude of the vectors[19] does not affect similarity. This allows text of different lengths to be compared. All that matters is that the proportions of the words in

[19]The lengths of the lines in figure 4.1 and a result of the size of the originating text.

the text remain similar between similar texts. Therefore, whilst the magnitudes of the vectors A and C are very similar[20] the cosine similarity between the two is only 0.6, the value when comparing B and C is 0.998[21] despite the lengths of the texts being significantly different[22].

4.2.4 Okapi BM25F

An alternative approach is BM25 (also known as Okapi BM25[27, 230] and shown in (4.13)). BM25 is a probabilistic ranking system which calculates the probability of document relevance based on the appearance of the query terms in the searched documents. Relevance is assumed to be a boolean property and so the relevance value calculated is in terms of chance not degree.

$$\sum_{t \in Q} \cdot \frac{\mathrm{tf}(t,d)}{k_1((1-b) + b\frac{l_d}{\mathrm{avg}(l_D)}) + \mathrm{tf}(t,d)} \tag{4.13}$$

Given a search query Q (consisting of keyword terms t) the probability of relevance of any document d can be calculated. $\mathrm{tf}(t,d)$ is the term frequency of keyword t in d. k_1 and b are free parameters (1.2 and 0.75 according to Billerbeck and Zobel [27]). $b \in [0,1]$ and controls the effect of document length on the results. l_d is the length of document d while $\mathrm{avg}(l_D)$ is the average document length of all documents in the search space.

However, given that records contain multiple separate fields (as opposed to documents where a single block of text is available) BM25F may be more suitable, this is a modified version of BM25 designed to operate on structured documents (e.g. eXtensible Markup Language (XML), JavaScript Object Notation (JSON)) [186]. The advantage of BM25F over BM25 for records is that it allows different fields within a single record to be weighted separately depending on their confirmed (or assumed) validity. For example, in the case of the ERPS records it may be desirable to give preference to those records which contain the search keywords in the *title* field, as opposed to those containing them in the *description*.

$$\mathrm{weight}(t,d) = \sum_{c \in d} \cdot \frac{\mathrm{tf}(t,c) \cdot w_c}{(1-b_c) + b_c \cdot \frac{l_c}{\mathrm{avg}(l_c)}} \tag{4.14}$$

[20] Both correspond to texts containing 12 words.

[21] $\cos(B,C) = \frac{(5\cdot 9)+(2\cdot 3)}{\sqrt{5^2+2^2}\cdot\sqrt{9^2+3^2}} = 0.998$

[22] 7 and 12 words respectively.

$$\sum_{t \in Q} \cdot \frac{\text{weight}(t,d)}{k_1 + \text{weight}(t,d))} \cdot \text{idf}(t) \tag{4.15}$$

Expanding on the BM25 notation, c represents a field in document d. w_c is a weighting applied to c. l_c is the length of field c and $\text{avg}(l_c)$ the average length of that field across D. b_c is the same as b from BM25 but varies per field (still in the range $[0, 1]$). First calculated is $\text{weight}(t, d)$ which is the effect of a single term t in the search query Q across all field c in the individual document d. These individual term values are combined with the $\text{idf}(t)$ to ensure that common terms do not overpower all others (i.e. *the*, *a*) and the values for all terms are combined.

4.2.5 Conclusions

Since the text similarity metrics discussed above are reliant on the same words appearing in both texts being compared, they are only effective when used against reasonably long pieces of text where there is a greater chance of the same words appearing. Against short pieces of text, such as those found in photo-history records, they are less effective[23]. This is the same issue which was discussed in section 2 with regards to query expansion.

Assuming that the issues surrounding the briefness of the record text can be resolved/mitigated, BM25F looks like an interesting way forward. The individual field weightings could potentially allow every text fields in a record[24] to be compared in one go.

4.3 Semantic string comparison

Semantic string comparison methods compare textual segments by identifying the similarity (or dissimilarity) of the concepts described in the text. Since it is the concepts that the text's terms describe that are being compared and not the terms themselves, semantic comparison techniques are far more resilient to the issues posed

[23] An analysis of 699,520 record titles, randomly selected using the British Museum (BM)[55] SPARQL endpoint[2], across 499 different artefact types (everything from blankets to shrines) found an average title length of just 4.81 words. An analysis of 8,446 photographs selected in the same way found an average length of 4.01. While photography records do have slightly shorter text, the briefness of GLAM records in general is also clear. The limitations of textual similarity approaches such as cosine similarity are therefore applicable, not just to photo-history, but to GLAM collections as a whole.

[24] I.e. *title* and *description* in ERPS records.

when compared with short pieces of text. This is because although the concepts described in the texts may be similar the terms used to do so are completely separate.

4.3.1 Latent Semantic Analysis (LSA)

LSA is a vector space model developed and patented in 1989[51] which uses a truncated document term matrix to discover underlying or 'latent' connections between terms in the documents being compared[116]. As an approach, LSA[25] is probably the most well known Short Text Semantic Similarity (STSS) method.

LSA uses a document term matrix (A) of size $m \cdot n$ where m is the number of documents and t is the sum total number of terms appearing in d documents. Each element of matrix A can be described as a_{ij}, i.e. the frequency of the i^{th} term in the j^{th} document. A weighting function is applied to all non-zero elements of A, typically TF-IDF, so as the reduce the importance of common, poor distinguishing terms[26]. The result is a sparse weighted document term matrix (X). The matrix is generally very sparse as each document will only contain a fraction of the t terms. A matrix of $\leq 1\%$ non-zero values is usual[116].

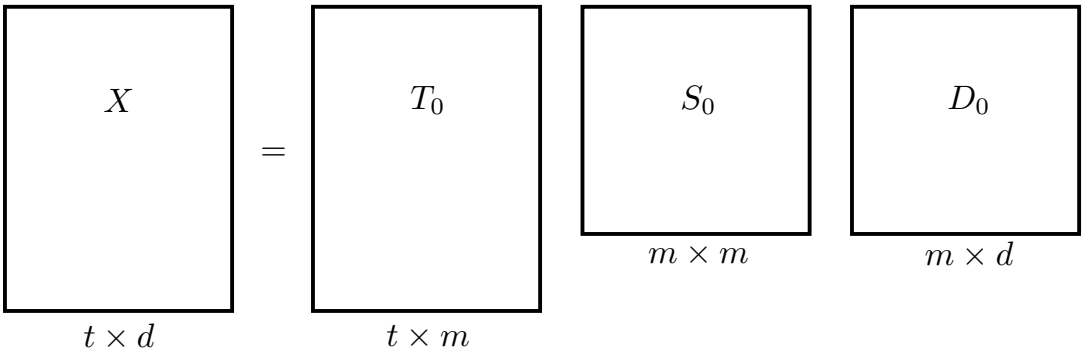


Fig. 4.2: Initial matrices used in LSA.

Singular Value Decomposition (SVD) is then used on the weighted document term matrix (X) to produce the orthogonal component matrices T_o, S_o and D_o as shown in figure 4.2 where $m \leq min(t, d)$. Redundant portions of the matrices can then be removed to create the new matrix $\widehat{X}$ where $\widehat{X} = TSD \approx X$. Redundant sections are removed by selecting the first k values as shown in figures 4.3 and 4.4.

Document similarity can then be calculated using a vectorial comparison of the vectors in $\widehat{X}$ using any of the techniques discussed in section 4.2.2 although cosine

[25] Also called Latent Semantic Indexing (LSI) when referring to informational retrieval.
[26] See section 4.2.1.

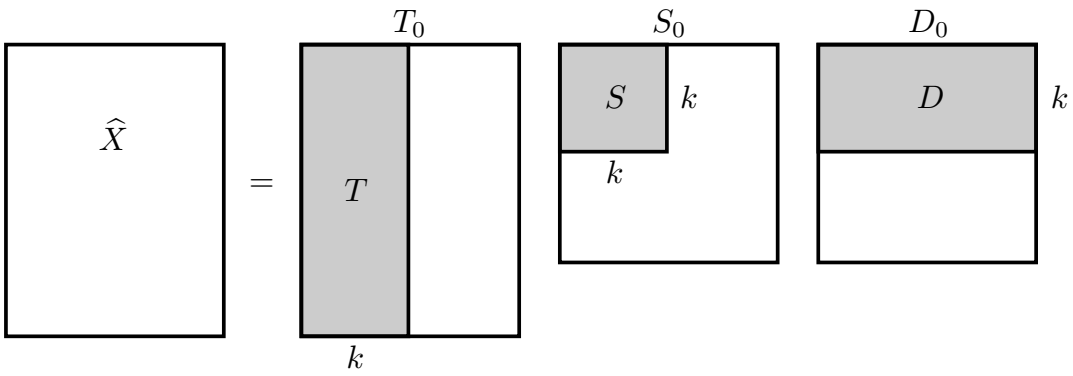


Fig. 4.3: Truncation of SVD matrices.

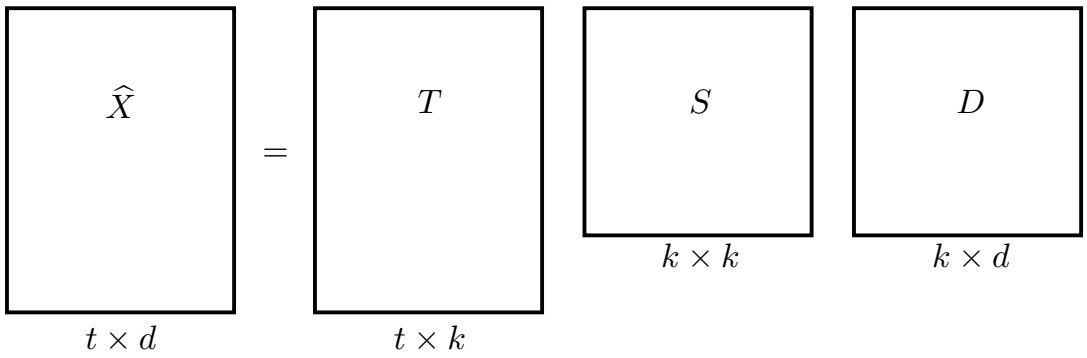


Fig. 4.4: Truncated/optimised matrices used in LSA.

similarity is more usual.

Importantly, LSA ignores the term order of the text and assumes that each term represents a single concept. This means that LSA can produce unpredictable results when faced with documents containing polysemy. This in turn means that LSA is best suited to comparing texts from within a single domain area where only one meaning of a term is likely to be used as opposed to comparing text across domains.

4.3.2 STASIS

Created for use in conversational software agents, STASIS uses a pre-existing LDB to identify term similarity. In contrast to LSA (amongst others), STASIS takes word order into account. This is a major distinguishing feature of STASIS when compared to other approaches which treat text as a 'bag of words'[73]. STASIS can be viewed as two separate similarity metrics joined together. One takes into account the semantic similarities between the terms and the other takes into account the word order similarity. The values from both of these are then combined to produce

the final similarity.

Overall semantic similarity is calculated as the cosine of two term vectors (see section 4.2.3). However the vectors are first modified according to semantic similarities between the contained terms and the importance assigned to each term. In the case of the work by Li et al.[123] these word similarity values are calculated using WordNet. However, STASIS itself just specifies a method for generating word similarity values across a hierarchically ordered knowledge base. Therefore whilst WordNet is the best known source for the similarity values, it is not the only option since any hierarchically organised source could be used.

The term similarity approach taken by STASIS is to combine the shortest path distance and the depth of the terms in the hierarchy. The reasons Equation 4.16 shows how this is achieved.

$$s(w_1, w_2) = \mathrm{e}^{\alpha l} \cdot f_2(h) \cdot \frac{\mathrm{e}^{\beta h} - \mathrm{e}^{-\beta h}}{\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}} \tag{4.16}$$

l represents the path length between two terms and h represents the depth of the subsumer (ancestor node) in the WordNet hierarchy. α and β are tuning values, both of which should have values $\in [0, 1]$. Li et al.[122] used values of $a = 0.2$ and $\beta = 0.45$ which represent the correct tuning when using WordNet.

Semantic similarity between the vectors is modelled as follows; for each term in the common vector T, if the term appears in the vector (T_1 or T_2) when set the value in the corresponding semantic vector to be 1. If the term does not appear in the vector then find the term in the vector with the highest term similarity. If the term similarity exceeds a pre-set threshold then set the value in the semantic vector to be the term similarity value. Otherwise set the value to be 0.

Individual term importance is determined using the information contents values taken from the Brown corpus[72]. The similarity values from the semantic modelling and the term importances are combined as shown in equation 4.17.

$$s_i = \tilde{s} \cdot I(w_i) \cdot I(\tilde{w}_i) \tag{4.17}$$

In a "bag of words" approach [a b c] = [c b a] despite the differences in element ordering. STASIS however includes a computational method for measuring word order similarity between texts and so will not consider the two equivalent.

The first step is to convert T_1 and T_2 into word order vectors (r_1 and r_2), this

requires assigning every term in the joint T set a numerical value. Converting term vectors to work order vectors is simple if the same terms appear in both T_1 and T_2. In the likely event that the texts being compared contain different terms, those terms that appear in only one vector are replaced with the numerical value of the T term that they have the highest similarity to, assuming that it exceeds a pre-set threshold. Those terms with similarities that do not exceed the threshold are replaced with 0.

$$\begin{aligned} T = T_1 \cup T_2 &= [\text{a b c}] & \\ T_1 &= [\text{a b c}] & \rightarrow r_1 = [1, 2, 3] \\ T_2 &= [\text{c b}] & \rightarrow r_2 = [0, 2, 1] \end{aligned}$$

A word order similarity value can then by generated by simply calculating the normalised difference of the word order vectors (see equation 4.18).

$$S_r = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} \tag{4.18}$$

The overall STASIS similarity for the pair of texts being compared given by equation 4.19. Where $\delta \leq 1$ and controls the relative effect that the semantic similarity and word order values have on the overall text similarity value. Li et al. state that δ should be kept at a value > 0.5 as word order plays a lesser role in text processing[123, 232].

$$S(T_1, T_2) = \delta S_s + (1 - \delta) S_r \tag{4.19}$$

4.4 Conclusions

LSA and STASIS demonstrate potential methods of identifying semantic similarity in short sections of text. Thanks to their ability to model semantic relationships between terms, they are significantly more suitable for processing GLAM records then the binary vector techniques, cosine similarity or Okpai BM25. However, the increased sophistication of the approaches comes at an increased computational cost

which may cause performance issues given the number of GLAM records available for examination. Despite this, the increased quality of the results promised by semantic similarity approaches cannot be ignored and so a semantic comparison metric is clearly required for comparing records from the ERPS collections, particularly for comparison of the *title* and *description* fields.

The presence of names in the data, specifically in the *person* field, means that other approximate textual similarity approaches such as Jaro and Jaro-Winkler are also very likely to play a part in handling the variation found in GLAM records.

5

Collections

Metadata is "data about data"[43]. Within the context of GLAM records and photo-history collections in particular, metadata typically means information about an underlying physical artefact. While metadata can refer to any information about an artefact (handwritten notes could qualify), within this section the focus is digital information held in a machine readable structure, i.e. XML or JSON. For this research metadata is seen as a means of resource discovery, storing useful information in computer readable structures which can be ingested and studied in order to identify relevant items.

Metadata has two roles in resource discovery, firstly it can represent non-interpretable information[1] in more accessible forms, i.e. textual descriptions of images[40]. Secondly an object's metadata can store additional information which is not present in/on the underlying object, i.e. a photographers name and date/-time/location an image was taken.

Object metadata in GLAM collections can also contain additional information such as an object's physical location, cataloguing information and/or conservatorial information. While valuable, this information is of limited use for this research and so will not be focused on.

The difficulty in querying a GLAM (or other collection) records is in understanding the format, structure and relationships of the metadata. This is not a major problem when manually searching. A person looking a specific collection record of a website can instinctively identify the information that they are interested in based on the field labels and the structure/content of the information. When it

[1] I.e. information that software cannot or that the specific software does not have the means to understand. For example, it is very difficult for software to examine an image and identify the image contents and so that information is effectively inaccessible.

comes to computerised information retrieval systems this instinctual knowledge and understanding is not available. Systems must be explicitly told where and what each piece of information is. There are several layers of additional structure that can be applied to metadata information in order to make it more understandable by software systems and these are discussed in the following sections.

5.1 Markup languages and metadata schemas

Markup languages and metadata schemas offer the minimum realistic level of structure for formatting data to be understood by software. Markup languages are a method of annotating documents in such a way that the annotations are identifiable. Examples include XML, Standard Generalized Markup Language (SGML) and JSON. These allow the structure of the metadata to be stored within/around the metadata.

While markup languages allow for the structuring of metadata, metadata schemas control how it is structured i.e. they state how the discrete pieces of metadata should be identified. For example, under the Dublin Core schema an items title should be marked with the identifier "title", it's creator should be marked "creator" etc[5]. As schemas are markup independent this can be represented in multiple markup languages as figures 5.1, 5.2 on the next page and 5.3 on the following page show.

```
1 <title>Example title</title>
2 <description>An example record</description>
3 <creator>David Croft</creator>
```

Fig. 5.1: Example of the Dublin Core schema marked up in XML.

One example of a popular GLAM metadata schema is Dublin Core. Dublin Core is a comparatively small schema, originally consisting of just fifteen elements[2] intended to be used to describe online resources such as websites[5, 43]. Unfortunately due to issues surrounding the trustworthiness of such self identified information[3] Dublin Core data is ignored by the major search engines[6]. It has,

[2]These now constitute Simple Dublin Core and are Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title and Type.

[3]I.e. metacrap, namely that people lie[1].

```xml
<?xml version="1.0"?>

<dc>
  <dc:title>Example title</dc:title>
  <dc:description>An example record</dc:description>
  <dc:creator>David Croft</dc:creator>
</dc>
```

Fig. 5.2: Typical example of the Dublin Core schema marked up in XML, following XML conventions and recommendations. The "dc:" prefix to the identifiers are used to mark the namespace of the identifiers, this potentially allows multiple schemas, with overlapping identifier names, to be utilised in a single XML document.

```json
{
  "title": "Example title",
  "description": "An example record",
  "creator": "David Croft"
}
```

Fig. 5.3: Example of the Dublin Core schema marked up in JSON.

however, seen adoption and adaptation[4]within more rigorously managed domains, i.e. GLAMs, where greater confidence can be had in the accuracy of the information.

Unfortunately there are many different schemas in use at different GLAM institutions. In 2013 the Canadian Heritage Information Network (CHIN) identified the existence of at least forty different metadata standards[5][9] and an analysis in 2011 of fifty four Europeana institutions found fifteen different formats[40]. This proliferation of different collection schemas has occurred because cross-collection searching was not considered as a major factor when the digitisations projects began. As discussed in section 1 on page 1, digital collections were initially created as a preservation and organisational tool and not for increasing collection access[8]. Since the information would only be accessed/used within an individual institution, the metadata schemas could be selected or created to fit the specific requirements or preferences of the individual institution. The change in collection focus to

[4]E.g. Europeana Semantic Elements (ESE), an expansion of the fifteen Simple Dublin Core elements with an additional thirteen tuned to the needs of the Europeana project[40]. See also Electronic Theses and Dissertations Metadata Standard (ETD-MS), Gateway to Educational Materials (GEM) and Rare Materials Descriptive Metadata (RMDM)[34].

[5]34 in museums, 6 in libraries.

increased accessibility[8] and cross collection searching resulted in these collections being made available online but still in their original schemas.

Although querying across different schemas is possible this can add significant complexity as the number of different schemas increases[34, 240]. A standard approach is to map differing datasets into a single schema before querying[34, 244]. Some schemas[6] have been specifically designed so as to assist in aggregating records from multiple sources and schemas[39, 202]. The mapping of one schema to another is often manual operation as it requires a high level understanding of the contents of the separate schema elements.

5.2 Ontologies

An alternative to manual mapping of metadata schemas is an ontological approach. A step above metadata schemas, while a schema means that metadata has a structure that can be understood by software, an ontology encodes that structure in a format that can itself be understood by software[7]. Importantly ontologies encode the meaning behind individual elements as well as the interconnections between the elements in a computer readable format. This allows for information stored in different schemas to be easily compared[128, 217]. Information stored in different schemas and labelled with different identifiers can then be compared automatically because they share the same semantic meaning as identified by the ontology. I.e. the knowledge that the identifier "hasAuthor" in one schema has the same meaning as the identifier "creator" in another can be represented in an ontology[40].
Relying on an ontology for schema mapping is, however, dependant on both schemas having a shared ontological framework. Within the GLAM community the standard ontology in use is the CIDOC Conceptual Reference Model (CIDOC-CRM)[3] but only a limited number of collection have for far incorporated this into their collection, i.e. BM[2, 55] and Europeana[62].

[6] I.e. Lightweight Information Describing Objects (LIDO), an expansion on the earlier Categories for the Description of Works of Art (CDWA Lite) schema.

[7] Using languages such as Web Ontology Language (OWL)

5.3 Syntax independence

Although ontological approaches can assist in searching across collection when they are present, these still do not fully solve the heterogeneous data problem that GLAM records suffer from.

Although metadata schemas such as Dublin[5] and LIDO[39] describe a specific way to lay out the records in terms of identifiers, they do not[8] specify internal field structure[34]. That is to say that while schemas describe the information container but do not describe the formatting to used on the information itself, such schemas are called syntax independent[40]. Although multiple collections may be using the same schema, because of syntax independence those collections may represent the same piece of information in multiple distinct formats. It is this syntax independence which makes GLAM collections such a challenging search space.

For example the date 3rd of May 1910 can be stored as '03051910', '3/5/10', '5/3/10', '19100503', '03/05/1910', 'May 3, 1910' etc. Different collections using different formats is an annoyance but not a major problem. The relevant software can simple be programmed with the correct formats to use for each collection. The problem is individual institutions using differing formats within their collections. In those cases the intended meaning of the information has to be inferred by clues in the information, a difficult and sometimes impossible task. Continuing with the previous example, a field containing 'May 3, 1910' cannot be interpreted as meaning anything other than the 3rd of May 1910. The meaning of a field containing '03/05/10', however, depends on whether the format used was DD/MM/YY[9], MM/DD/YY[10] or YY/MM/DD[11] and in the case of historical records[12], it is not even clear which century is being referenced[13]. Without a known format, it is impossible to be certain that the field contents have been interpreted correctly.

[8] Or did not initially.

[9] As would be expected in most of Europe, the Americas, North Africa and Oceania.

[10] As would be expected in the USA.

[11] As would be expected in East Asia and a small portion of Europe.

[12] Such as those contained in GLAM collections.

[13] See section 6.3.6 on page 115 for a fuller description of the problems posed by *date* fields in GLAM collections.

5.4 Resource access

Regardless of the issues in interpreting metadata, in order to use it, it is first necessary to get it. There are several factors which need to be considered. First there is the method of accessing the information, just because information is stored digitally does not mean that it is accessible. Collection information can be stored in a variety of different formats, everything from simple text files and spreadsheets[14] to databases. While simple text files published on a website would technically count as having put collection information online, the usefulness from the standpoint of most website visitors would be limited. Databases are the more common method for storing the information although most individuals accessing the information will not access the database directly. Intermediate layers such are websites and Application Programming Interfaces (APIs) will provide restricted access to the underlying, database held, information. Some collections websites are be built on top of their collection API[15], querying the underlying database/s through the API and converting and formatting the results into a suitable, human friendly format.

While it may be possible to gain access to collection information held in flat files or complete copies of databases[16], this chapter will be looking at collection access via APIs. This is because collection APIs have been specifically designed to provide access to the record collections in a software friendly manner.

The two predominant designs for collection APIs are REpresentational State Transfer (REST) and SPARQL interfaces. REST interfaces are the more common and varied of the two approaches, whilst SPARQL interfaces are more capable. The level of variation within REST interfaces is due to the fact that REST refers to a series of informal best practises, rather that a laid out specification. As the name may suggest, SPARQL offers similar capabilities to that of Structured Query Language (SQL) which is the standard method for querying information held in relational databases. SPARQL operates on RDF structured data[17], it was designed to perform complex joins across multiple disparate datasets. The SQL like mature of the language provides near unlimited freedom when it comes to querying the data as any combination of fields and factors which can be described in SPARQL can be combined in a single query. This is a significant advantage over REST interfaces which

[14] E.g. flat files, Excel etc.
[15] I.e. the Victoria and Albert Museum (V&A)
[16] Database dumps.
[17] Which can be visualised as a series of three column tables from a SQL standpoint.

only provide those search options and combinations that the interface's designer intended.

Both types of interface will return their results as either XML or JSON, some interfaces will offer both[18]. In the case of SPARQL interfaces the results will be structured RDF. For REST interfaces the record ontology will depend on the individual collection. As can be seen in section A on page 190, every REST interface consulted during this research project returned their results in a different schema.

5.5 Conclusion

The major challenge facing any attempt to search across multiple GLAM collections is one of metadata schemas differing across multiple institutions. While ontological approaches may

Moving all GLAM institutions onto a single metadata schema would be the preferred approach from the standpoint of writing software[34]. The reality is that differing institutions have differing requirements and attitudes towards their collection data and a universal schema is unlikely. The excessive proliferation of differing schemas is not, however, likely to continue thanks to the establishment of shared schemas such as Dublin Core and knowledge of the problems that schema proliferation has caused. It is more likely that a small number of shared schemas will become more widely adopted.

What does not, however, appear to have been sufficiently addressed is the heterogeneous nature of the data within the fields. Going forwards, digitisation projects are likely to adopt stricter syntax standards as the problems of syntax independence become increasingly apparent, what is to be done with already digitised records is not clear. The solution to this problem is not a simple one, information that contains randomly formatted information is not easily automatically converted into another. Conversion can be done manually but given the current size of GLAM collections, the cost of digitising them once[174] and the backlog of undigitised records[173, 204], it seems unlikely that existing records will be re-digitised any time in the near future.

[18] E.g. the BM SPARQL and the V&A REST interfaces.

6

Methodology

This chapter describes the methods and techniques that were used/developed throughout this research to identify co-referent records. Discussed is not only the comparison of the individual fields but also the overall steps and sequence which takes a single a starting record through to the final co-reference analysis. Also covered are the failed attempts and dead ends where it is believed that such knowledge is necessary to understand the methods that were finally settled upon.

The flow of processes from start record to final analysis is complicated and has multiple parallel streams, discussion of the methods will, therefore, be ordered chronologically. The layout of the individual sections and subsections with relation to the overall flow is shown in figure 6.1 on the following page.

To summarise the proposed approach and as figure 6.1 illustrates, the initial seed record[1] is computationally examined to identify the keywords that it contains[2].

Those keywords are then fed into a query expansion system[3] in order to produce a list of search terms containing, not only the original keywords, but also semantically similar terms and inflected forms of those words. The expanded keyword list is then used to search multiple external GLAM collections[4] via their APIs[5] for any records which contain at least one of the terms in the query expanded keyword list. This can potentially be tens of thousands of records.

With the records collected from the external collections, the field names and

[1] A specific ERPS record that is being searched for in the external collections.
[2] See section 6.1 on page 91.
[3] See section 6.1 on page 91.
[4] See section 6.2 on page 93.
[5] An Application Programming Interface (API) is simply a set of routines, functions and/or protocols which describe how a piece of software can and should interact with the software, library or website providing the API.

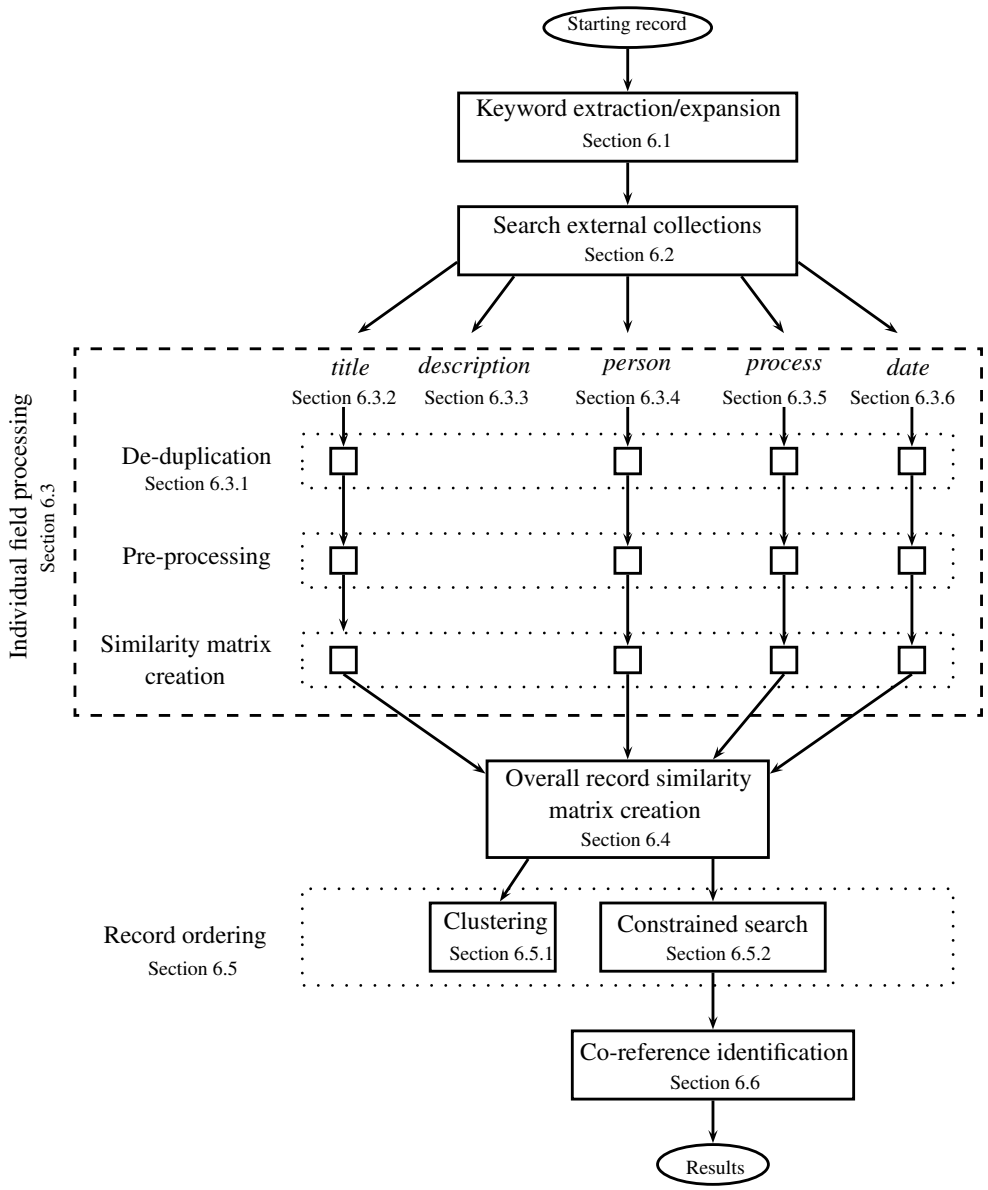


Fig. 6.1: Process flow diagram for the proposed approach.

layout of the records are normalised[6]. Certain fields[7] are de-duplicated and the field contents processed into standard formats using specially created pre-processing algorithms. A non-directional pair-wise comparison of the de-duplicated and standardised fields is performed to produce a similarity matrix for each field using a custom similarity metric designed for each field. The individual field similarity metrics are then used by a custom FIS to produce an overall record similarity matrix.

Finally, the overall record similarity matrix is processed using a dendrogram

[6]See section A on page 190.

[7]*title*, *person*, *process* and *date*.

generation algorithm in order to sort the records and identify those with the highest similarity to the seed record. Those records which appear closest to the seed record in the dendrogram, are the proposed approach's best guesses for valid co-reference candidate records.

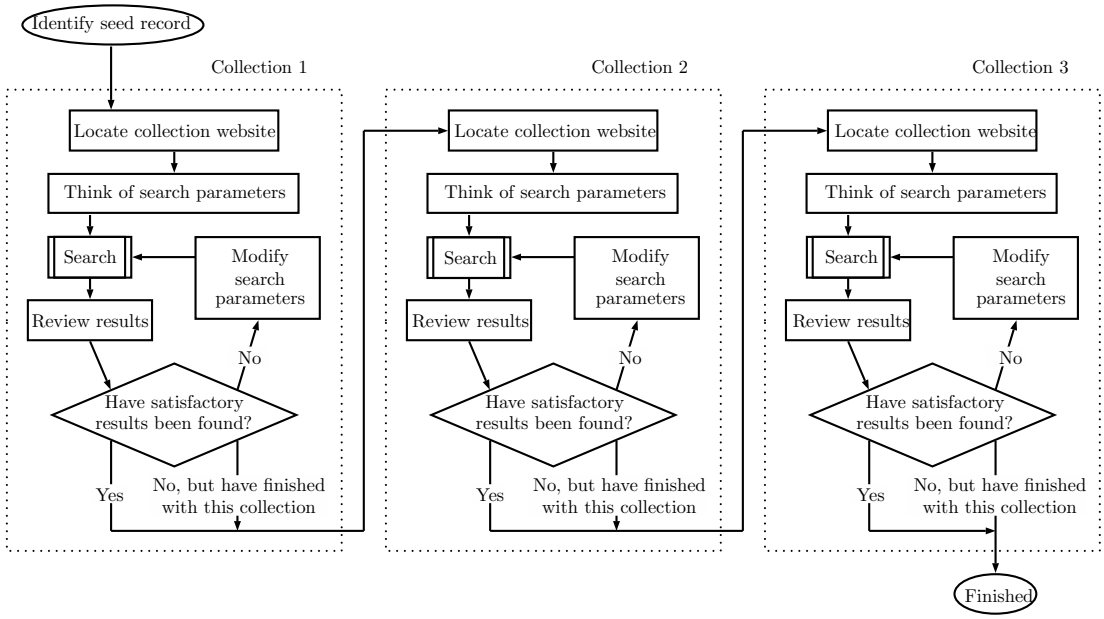


Fig. 6.2: The actions needed in order to search through three collections using traditional search interfaces.

The focus of this research was to determine if it is possible to automatically identify high quality co-reference matches across digital humanities collections at all. By automatically performing or bypassing many of the steps required in searching using traditional search interfaces, a time saving in terms of person hours can be achieved. A demonstration for why the proposed approach would save individuals time when searching can be seen in figures 6.2 and 6.3 on the following page. As figure 6.3 on the next page shows, many of the actions needed for manual searching are bypassed under the proposed approach. By reducing the number of actions that need to be performed, the amount of time that an individual spends searching is also reduced. Figure 6.2 also demonstrates that when searching across multiple collections, some actions needed to be repeated at each additional collection. As the number of collections being searched increases, so does the number of steps. In comparison the number of user actions required under the proposed approach remains static regardless of the number of collections being searched.

The actual time savings produced by the proposed approach can not be presented

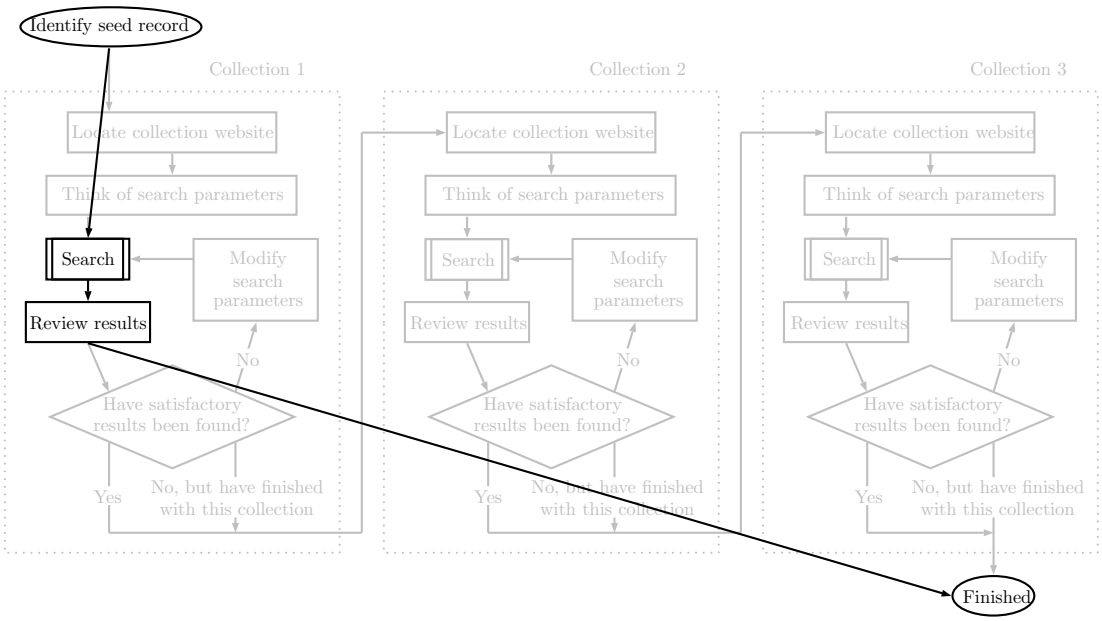


Fig. 6.3: Manually performed actions to search through three collections using the proposed approach. Overlain on the diagram seen in figure 6.2 on the preceding page.

at this time. The software implementation used during this research broke down the proposed approach into a number of separate programs. While this approach had advantages in that it allowed for easy diagnosis of problems, it has a major disadvantage in that it is not computationally efficient. The time that the proposed approach currently takes to calculate co-reference candidates does not, therefore, bear any resemblance to the time that a real world deployment of the proposed approach would take. The precise time savings of the proposed approach were not, therefore, part of this thesis. If the proposed approach is shown to be capable of identifying high quality co-referent record matches, then future work would involve rewriting the source code into a single, computationally efficient process. This would not, however, have been a sensible use of resources until the proposed approach has proven the quality of its record matching.

6.0.1 Lack of image processing

Given the photographic nature of the records being investigated during this research, image processing techniques might appear to be a sensible area of investigation. In particular, techniques such as Speeded Up Robust Features (SURF)[20], Scale Invariant Feature Transform (SIFT)[130] and Rotation Invariant Feature Transform

(RIFT)[117] which are capable to identifying matching images even when those images are rotated, partially obscured or otherwise transformed. There are, however, several issues which complicate the use of image processing techniques.

Firstly, retrieving the image for each record in addition to the record metadata would significantly increase the time needed to query the external collections. A single image is many times larger than textual metadata which would mean a longer download time but the primary issue is the number of additional API calls that would be required. For all the collection APIs looked at in this research[8], accessing a record's image would require at least one API call per record. Combined with restrictions on the rate of and number of API calls laid out in the API terms and conditions[9], the number of additional API calls would have very significantly increased the time needed to query the external collections.

Secondly, image processing is computationally expensive. Comparing images would be expected to significantly increase the processing time of any solution. This applies even with relatively fast algorithms such as SURF[20]. These factors would not, however, have been insurmountable problems. One possible way forwards would have been to use an analysis of the textual metadata to reduce the size of the search space, with image processing being performed of a much smaller subset of the overall search space[10].

The main barrier though is the lack of images. As discussed in section 1 on page 1, 97% of the ERPS collection lacks any image data. The 3% remaining contains a number of sketches of the original photographs which have limited detail and questionable accuracy, both of which which make any image processing significantly more difficult.

In the end, the near total absence of ERPS images that could be processed meant that investigating or experimenting with image processing algorithms was not seen as a sensible way forwards.

6.1 Keyword extraction/expansion

As discussed in section 2, the quality of the results produced by a search system can be measured in two distinct ways. Firstly there is the precision, the number of relevant results returned compared to the number of irrelevant results. Secondly

[8] See section A on page 190.
[9] Namely those of the V&A.
[10] A blocking/chunking approach[19].

there is the recall of the system, the number of relevant results as a ratio of the number of relevant results in the search space as a whole. Whilst a recall of a 100% can easily be achieved by simply returning every record in the search space, this does tend to defeat the purpose of a search system as examining and processing the entire search space is not a realistic option. The combined size of the collections that were used for this research is tens of millions of records (see table 6.1). Attempting to compare the ERPS records to every single photographic record available online in even a single collection is inefficient and impractical. Instead, for this research a sub-section of the overall search space is selected which can then be subjected to a more intense scrutiny (chunking/blocking the search space). This is achieved through a combination of keywords extracted from the seed record[11] and query expansion. Keywords from the seed record are used as this makes it possible to identify those records which bear some resemblance to the seed record. The aim is to select from the search space all the records which could be considered to have any resemblance to the seed record. For this discussion "any resemblance" is defined as containing at least one of the keywords. Keywords are not ideal variables for blocking/chunking the records[76], but given the limited nature of the external collection APIs and the search restrictions those enforce, keywords are the best option available. However using only those words which were directly collected from the seed record could still exclude a large number of co-referent records.

Due to the briefness of the record text and the large number of distinct terms that could be used to describe each record, it was determined that selecting the subset records based solely on the keywords in the seed record would be ineffective and so some form of query expansion was required. Whilst relevance and pseudo-relevance feedback approaches have shown the best results according to the existing literature (see section 2.3.2), given the briefness of the text it was not felt that they would be effective here. An added concern is the iterative nature of relevance feedback techniques. These may be suitable when querying local collections, but repeated calls to an external collection would be prohibitively time consuming. Therefore, it was decided to use WordNet as the LDB for a global reference approach (see section 2.3.1.)

The global reference approach used is the simplest method available. Having

[11]The term "seed record" refers to a single record which co-reference matches are being sought for. For this research this would be one of the 1,040 ERPS records with image data.

generated the list of possible synsets[12] for each of the initial keywords, the list of synsets is expanded to include the top three hypernyms and holonyms for each of the top three synsets. This gives a list of $n + 3n + (3 * 3n) + (3 * 3n) \rightarrow 22n$ potential synsets where n is the number of original keywords. The lemmas[13] for all the synsets are then added as new, expanded keywords.

It would have been preferable not to have needed to implement a limit to the number of synsets, hypernyms and holonyms etc. This limited approach means that only the most statistically likely synsets get expanded. If the keyword being expanded does, in fact, correspond to a comparatively uncommon synset, then valuable terms will be missed. However without these limits the number of terms becomes unmanageable, simply because of the number of records returned from the GLAM collections and the amount of time which in required to collect and process them[14].

6.2 Searching external collections

Once the list of keywords to search for is available, querying of the external collections can begin. As discussed is section 1, the relatively recent shift in GLAM institution focus[8] means that more and more GLAM collections are being made available online. Whilst it is possible to access these collections via their traditional human usable websites or by requesting copies of collections, these approaches would have required a significant amount of time and resources to be invested in the development of screen scraping systems and/or negations for each collection that was used. In both cases there are also potential legal issues (i.e. copyright)[7] and negotiations which would have needed to have been addressed.

Neither the time and resource requires, not the potential legal issues would have presented an insurmountable problem had there been a compelling reason to do so. As it was, the decision was made to only consider those collections which:

1. Are openly available online. This circumvents the problem of gaining permission to access the collections, as long as any relevant terms and conditions are adhered to (e.g. number of queries allowed per hour/day etc). This means

[12] Synsets are a collection of semantically similar or identical items.

[13] The words associated with a synset. For example synset ball.n.03 in WordNet has the lemmas 'ball', 'globe' and 'orb'.

[14] See section 7.1.3 and in particular 7.1.3.1 for evidence of the issues posed by excessive query expansion.

easy access to a large pool of potential data.

2. Have REST or SPARQL interfaces. This dramatically reduces the problem of accessing the records in a suitable format. Whilst more and more collections are being made available using these formats, at the present time the numbers lag noticeably behind those of traditional websites. However, for the purposes of this research the number of available collections is more than adequate although this issue would need to be revisited if wider searches were attempted.

The decision restrict the collections considered in the manner had added advantages. Namely that it forced the research to face the very real restrictions and limitations of the collection interfaces while also providing hands on experience with those same interfaces.

All of the institution and APIs offer keyword searching of their collections. However depending on the institution and the sophistication of the API, other search options and filters can be available[15]. When additional filters were available, they were not used for two reasons. Firstly, the additional filters removed records with empty fields and which did not, therefore, match the filters[16]. This was an undesirable behaviour. Secondly, the additional filters were only available for some of the collection APIs and so would have complicated the record collection process and introduced additional failure points. For example, if using date filters the *date* field values could be misinterpreted, potentially causing valid records to be excluded from the results.

6.2.1 Simulating collection APIs

During this research, for both the initial experimentation and subsequent testing, real GLAM APIs were avoided whenever possible. Instead, more than 1.7 million[17] records were collected from several GLAM collections and stored in a local database. Experimentation and testing was then conducted against these records, as opposed to using the original online collections. The reasons for this decision were two fold. Firstly, this removed the dependence on the collection APIs being available at all times. Secondly, querying the local files instead of the real APIs prevented

[15]For example filters based on dates or location possible with the V&A API.

[16]E.g. if filtering for records from after a certain date, records with no date information would be excluded.

[17]1,783,278.

the queries from either swamping the API servers with hundreds of calls to the detriment of other users or needing to reduce the rate of querying and potentially waiting hours for results to be returned. However, efforts were made to replicate the behaviours of the collections APIs as much as possible so as to avoiding ending up with a search approach that could only be used against locally held records.

In order to simulate better the restrictions of the real collections, access to the collections was only allowed via the same keyword based searching that was available from the collections' APIs. Therefore whilst local copies of the records were available and any search technique that was desired could have been used, the approach presented in this thesis is designed to operate within the restrictions resulting from using external collection APIs. The sole difference was that the limits on the rate or number of calls were not included.

Ideally full copies of entire collections would have been made. However, the interfaces as they are configured simply do not allow it. Instead, as large a subset of the collections as possible was downloaded given the restrictions of the APIs .

Brute forcing a complete set of the records from the collections using a comprehensive list of dictionary words was briefly considered, but ruled out for reasons of time and likely violations of the terms and conditions of the collections. Instead, since the aim was to locate co-reference matches for the 1,040 ERPS records with visual information, only those records which were relevant for that task were targeted. A list of all (valuable) words listed in those records (both *title* and *description* fields) was generated. This had an added advantage over using a dictionary as the generated list included person and place names not found in a dictionary. This list was expanded with the synonyms supplied by WordNet to produce a list of 3,846 words, which were used to query the external GLAM collections via their APIs. Over a period of several weeks[18], this list was used to query the Brooklyn Museum (BkM), DigitalNZ (New Zealand) (DNZ), the Library of Congress (LoC) and the Victoria and Albert Museum (V&A) collections and the resulting records collected. All four of these collections use REST interfaces. The SPARQL interface of the British Museum was intended to be used, but unfortunately it had to be excluded

[18]It was necessary to adhere to the terms and conditions of the collections at all times. For example the V&A interface places a limit of three thousand API calls per day at no more than one per second which severely restricted the collection rate.

due to technical difficulties[19].

Querying the collection REST interfaces was done using a distributed system. Friends, family and co-workers were all recruited and asked to run a small program on their computer/s. This program downloaded search terms from a central repository and used those terms to query the collection APIs. The results were then uploaded back to the central repository. As this was not the same person querying the APIs, each instance of the software could query up to the maximum allowed daily queries without technically violating the terms and conditions for the various APIs. This massively increased the number of API calls that could be performed per day compared to a single individual.

In the distributed program API calls to the various institutions were interleaved, i.e. while the program was waiting to be allowed to query one collection, it could be querying the others. This meant that the software was able to query collections almost continuously whilst still following the terms and conditions regarding the rate of queries[20]. Querying the collection APIs in this manner took almost the whole of January 2012. Exact dates are not available as some sets of results turned out to be corrupted and had to be rerun at a later date. Also the number of running instances of the software varied on a day to day basis which strongly affected the number of results collected per day. The majority of the record collection, however, took place between 6/1/2012 and 29/1/2012, at least twenty separate computers were involved.

In the end 23,881,009 records were collected which, after further processing to remove duplicates[21], became 1,761,785 distinct records. The ERPS and Photographic Exhibitions in Britain (PEiB) collections added a further 21,493 records to the local database. As these collections were hosted by DMU it was possible to get full copies of the relevant databases from the Photographic History Research Centre[222]. Details of the various collection APIs can be found in section A on page 190 and information on the overall sizes of the collections and the number of results collected from each one can be see in table 6.1 on page 98.
Since all of the collections use and return results in different layouts and schemas,

[19]The interface was in Beta testing during this research project although it was openly accessible. The syntax needed for keyword searching with multiple search terms was not known until February 2012 by which point all other record collection was completed. As a large sample of records had already been collected, the decision was made not to spend another month querying APIs just to get the British Museum records.

[20]I.e. no more than one per second in the case of the V&A

[21]Individual records could be returned by multiple keywords. Duplicates within the records collected from individual collections were identified using the record UIDs.

the raw records collected had to be transformed into a single standard format. It must be made clear however, that the contents of the collected fields were left unchanged. The only changes were restricted to normalising the field names used[22].

The synonym expanded word list was not the only filter in place whilst searching. Several collections (e.g. V&A) restrict the number of results returned per query and have organised their records into sub-collections focused on particular areas (e.g. sculptures, paintings, photographs). When possible, the search queries were constructed so as to collect only relevant records (i.e. only the photography sub-collections were searched). Relevant records in this case were records pertaining to photographic negatives, positives, prints etc. and for which digitised copies of the collection item existed. It is possible that valuable records could have been excluded from the results by this digitised copy requirement. However, in order to determine if two records are, in fact, co-referent it is necessary to be able to compare the actual images to which the records are referring. Without digitised copies of the collection items this would most likely require physically travelling to the collections in order to compare them. Alternatively and depending on the collection, it is sometimes possible to request copies of, as yet, non-digitised items. While this might be acceptable if a small number of items are required, requesting that additional records be digitised or tracking them down physically for this research was not seriously considered.

There were two main reasons for this, firstly this research is intended to show that improved search techniques are possible, not to actually conduct a full scale search for the 'missing' ERPS images. In order to simplify the demonstration process, the test records from the ERPS records were restricted to those some form of image data so as to make it easier to see if the matches suggested by the new approach were in fact matches. Allowing image-less records from external collections to be included in the results would have significantly increased the time needed to examine the test results. Not just for the final testing, but also, more importantly, during the tens of tests required during the development and tuning of the software/algorithms where it would have had a much greater effect on the time required. Secondly, even with the record restrictions in place, $\approx$1.7 million records were collected. More records simply were not required to demonstrate the effectiveness of the proposed approach (or lack thereof). As such, including these image-less records would have

[22]See section A on page 190.

Collection name	Overall collection size	Records collected
Brooklyn Museum (BkM)		2,352
DigitalNZ (New Zealand) (DNZ)	>6,400,000	859,412
Exhibitions of the Royal Photographic Society (ERPS)	34,197	1,040
Library of Congress (LoC)	>13,300,000	875,267
Photographic Exhibitions in Britain (PEiB)	20,453	20,453
Victoria and Albert Museum (V&A)	>1,000,000	24,754

Table 6.1: Estimates of overall collection size and number of records actually collected for this research. API details for the collections can be found in section A on page 190.

cost significant time and effort, both in terms of this research but also for the individuals responding to the digitisation requests, for no obvious benefit. Whilst physically visiting image-less records or ask for copies would be a sensible step during an actual search, due to the demonstrative nature of this research it was unnecessary and counter productive. Therefore whilst filtering the records so as to exclude those without an associated image could have excluded potential co-reference matches for the 1,040 ERPS records, it would not have been possible to show that matches to those records were co-referent. For this reason, they were excluded.

6.3 Individual field processing

As the individual fields available from the ERPS records are very different from each other[23], a similarity metric which works for one will/may not work for another. Therefore, distinct metrics tuned to the particular challenges of each field were needed. Since several metrics were employed, each field was processed separately in order to produce a similarity matrix for that individual field.

6.3.1 De-duplication of field values

In order to improve the speed of the similarity matrix creation for the individual fields, the values of the individual fields were first processed in order to remove duplicates. This did not always produce a significant effect, but on fields such as

[23]I.e. different fields contain different pieces of information in very different formats and with very different meanings.

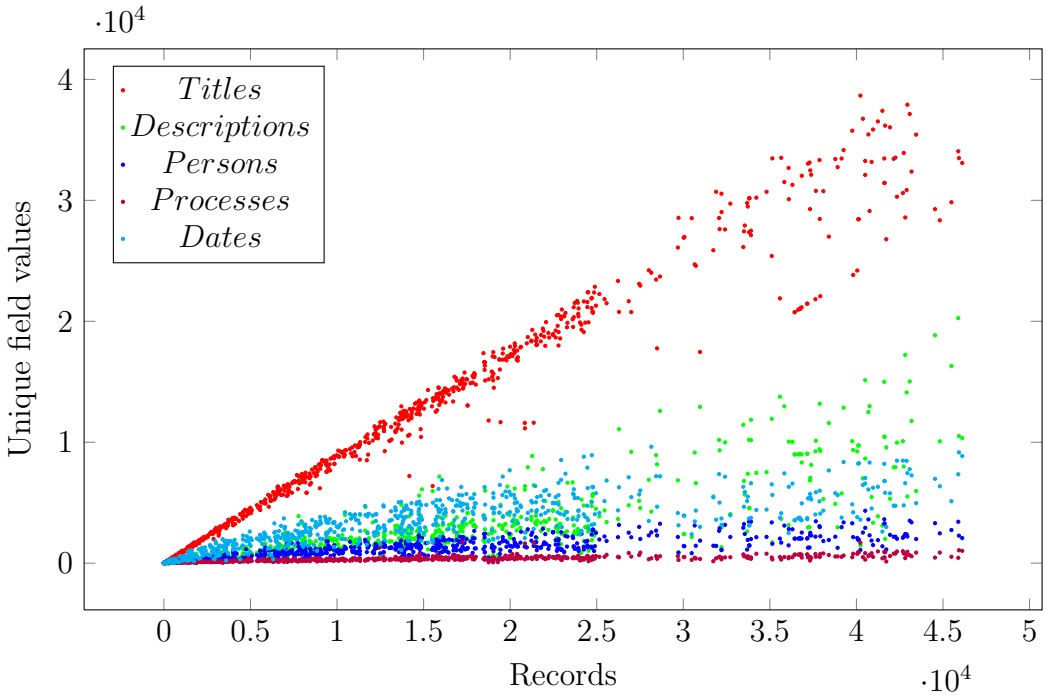


Fig. 6.4: Comparison of record counts from ERPS seeded searches versus the number of unique values per field.

process and *date* where a large number of fields are empty, the number of unique field values is significantly smaller than the overall number of fields being compared. The individual field values were compared on the basis of a character level string comparison. This is the simplest possible method of comparing the individual field values and, as such, the number of unique values could definitely be reduced further if the field values were compared using a more intelligent method.

However, even the string comparison method used produced a significant reduction in the number of the *description*, *person*, *process* and *date* values to be processed and a small but noticeable reduction in the number of *title* values as shown by figure 6.4. This in turn produced significant savings in the sizes of the similarity matrices for the fields[24].

[24]Taking for example the *person* metric. As shown in figure 6.4 and discussed further in section 7.1.3 on page 146, on average the number of unique *person* field values will be just 8% of the total number of records. Assuming 10,000 records then matrices for the deduplicated and non-deduplicated values would contain 319,600 and 49,995,000 values respectively. Or to put it another way, a 92% reduction in the number of values produces a 99.4% reduction in matrix size.

6.3.2 *Title* field metric

The *title* field is typically a very short description of the contents of the relevant photograph[25]. It may also be an emotional or artistic description of the contents[26].

The average number of 'useful' *title* words per record is very low[27]. Unfortunately, since the field rarely contains full sentences, the use of NLP is also difficult. The briefness of the text also excludes the use of standard approaches for measuring textual similarity (such as TF) in any effective way as even when two *titles* are semantically similar, the briefness of the text means that they are unlikely to contain any of the same words. With no words in common between *title* pairs, the term vectors will be perpendicular to each other in the term vector space and so the cosine similarity will be zero[89, 142]. Since statistical term count approaches were unsuitable, approaches that considered the semantic meaning of the field text, and which could, therefore, cope even in the absence of the same words appearing in *title* pairs, had to be considered.

Approaches such as LSA or STASIS could have been used at this point. However, LSA has a very high computational cost which would have caused significant problems when working on the data of this size[28]. Whilst STASIS's computational cost is much lower than that of LSA, its costs are still considerable[29].

Therefore, since neither the established statistical and semantic approaches were appropriate, a new, novel short text similarity technique was created which took into account semantic similarity between terms but which was also intended to sacrifice some modelling accuracy for the sake of significant reductions in computational complexity and, therefore, time.

In this novel approach, called and published[30] as Lightweight Semantic Similarity (LSS), semantic term similarities (semantic similarity between individual terms) are combined with vector similarity methods more typically usually used in statistical analysis. The performance of this new approach compared to existing techniques, in terms of both the accuracy of the similarity values produced and the computational cost, was tested experimentally and the results are included in the testing chapter[31].

[25] E.g. "The Entrance from the Cloisters, Canterbury Cathedral", erps16243.

[26] E.g. "Simplicity" and "A Labour of Love ", erps16207 and erps16254 respectively.

[27] Combining the *title* and *description* fields for the 34,197 ERPS records gives a mean average of 8.1 words. Filtering out low values terms (i.e. 'in', 'and' etc.) produces a mean of 5.4 words per record.

[28] See section 7.2 on page 151 for proof of this.

[29] See section 7.2 on page 151.

[30] See section sec:published.

[31] See section 7.2 on page 151.

6.3.2.1 Pre-processing

The original raw fields, which were collected from the external GLAM collections, contained information in a variety of different formats. This was a more obvious issue for the *date* field but it was present in varying degrees for all of the fields including *title*. Therefore before the fields could be compared by the various similarity metrics, the raw field values had to be converted into standardised representations as part of a process known as pre-processing.

Pre-processing was also used to reduce the computational load of the similarity metrics. This way the raw fields only needed to be converted into a standardised form once, instead of every time a field needed to be compared.

The first step is to generate a term vector for each *title* field. This involves cleaning and tokenising each *title*. Tokenising[32] means breaking up a piece of text into discrete chunks called tokens. Typically each token represents a single word and, as such, text is often tokenised by using white space or punctuation in the text to identify word boundaries. For example, the text "The Chrysanthemum Lady"[33] when tokenised would produce ['the', 'chrysanthemum', 'lady'].

Second, irrelevant and/or common terms are excluded[34]. Whilst many of the terms are common to any search (e.g. 'and', 'a', 'on' etc), some are specific to searches for photographs. For example, 'photograph' appears in 4% of the records collected[35]. As such, it is a poor identifier for distinguishing between records and is, therefore, included in the list of words to exclude.

Finally, the words are run through WordNet to identify the word's synsets (when relevant synsets exist). This stage also has the effect of normalising multiple forms of the same word (i.e. plural, past, present, future tenses) into a single representation which simplifies the comparisons at the cost of a small degree of precision. When synsets are not available, the words are in their raw form and compared using character based string matching. Removing words that lack synsets is not an option as these include many person and place names as well as some valuable technical terminology.

[32] Or tokenisation.

[33] erps17093.

[34] The list of words to exclude based off the nltk.corpus.stopwords list provided by the Python Natural Language Toolkit (NLTK) library[28]. The only additions were the terms "photograph" and "photographs".

[35] 65,491 of 1,783,280 records.

6.3.2.2 Similarity metric

The pre-processing produces produces two term vectors representing the words and number of occurrences of those words for the two *title* fields being compared. Also produced is a list of the corresponding synsets for each word.

This approach uses the cosine similarity of the two vectors as the similarity measure for the *title* fields being compared. Whilst cosine similarities of term vectors is a common approach for identifying document similarity[36], in those situations the term vectors in question are hundreds if not thousands of elements long. The briefness of the *title* fields means that is it unlikely that there will be any shared terms between pairs of *titles* even when they are semantically similar and, therefore, the cosine similarity of the vectors will be zero.

The novel aspect of this approach is the manner in which the initial term vectors are modified using the synset similarity values taken from WordNet. By calculating the cosine similarity on the weighted term vectors rather than the initial ones, it is possible to compare according to a pseudo-semantic similarity of the terms and so mitigate the issues caused by the briefness of the text.

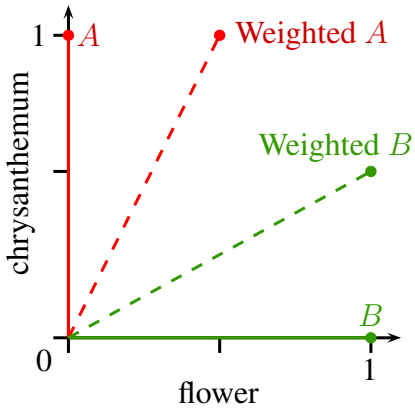


Fig. 6.5: Example of the difference between term vectors and weighted vectors.

A very simple example of this process is shown in figure 6.5. The example consists of two pieces of texts A = 'chrysanthemum', B = 'flower' and a similarity where sim('chrysanthemum', 'flower') = 0.5. Clearly the cosine of the original vectors (marked A and B in fig.6.5) would be 0 as the vectors are perpendicular to each other. Compare this to the result produced from the weighted vectors of the same pieces of text where the cosine similarity is 0.8.

[36] See section 4.2.3 on page 72.

Weighting the term vectors in this manner requires similarity values for each pair of synsets across both vectors. Since there are multiple synsets associated with each term, there are several possible similarity values which could be used depending on which pair of synsets are compared or how the results from multiple pair comparisons are combined. During this research the following approaches were tried.

1. First synsets - WordNet synsets are ordered according to their frequencies of appearance within the corpus which was used to produce WordNet[37][66]. Therefore, the position of a synset within the list of synsets associated with each term correlates to the likelihood that it is the intended meaning of the searched for word. There is one significant issue with this approach, just because the first synset is a more frequent usage than the second synset does not mean that it is significantly more common. Therefore, this approach frequently compares the wrong synsets and produces low similarity values even when the actual similarity is high.

2. Averaged synsets - The average of every synset in set A compared to every one in set B. As a result, this approach requires $|A| \cdot |B|$ comparisons. However even when the terms being compared are valid synonyms of each other the average similarity produced via this method is still low. The few high similarity comparisons were overwhelmed by low value comparisons to produce an overall similarity value that is slightly, but not significantly, higher than the values produced from a comparison of two random terms.

3. Conditional synsets - This approach produced the best results of all the techniques tried. Unfortunately whilst effective at comparing individual *title* fields, it proved too computationally expensive when used to generate similarity matrices. This approach identifies the synset to use for each term in vectors by calculating the best possible combination of synsets for each term vector as a whole. Assuming that each term can only have a single synset representing it, this approach tests multiple combinations of synsets in order to maximise the sum of path similarity scores. This means that if related words (as identified in WordNet) appear in a *title* field, then the synsets which reinforce each other are selected and different synsets can be selected for each term in different vectors.

[37] See section 2.3.1 on page 23.

4. Maximum synsets - Similar to the averaged synsets approach, this method simply takes the maximum path similarity value produced from the pair-wise comparisons. Whilst the results using this method were not as good as that achieved using the conditional technique at the level of individual pairs of *titles*, it is significantly less computationally expensive and, therefore, scales acceptably when used in similarity matrix generation.

6.3.2.3 Worked example

In order to properly describe the *title* metric, a worked example of a single *title* pair is included. In this example, the two *title* fields are defined as A and B, with the contents "the chrysanthemum lady" and "a woman selling flowers" respectively. Whilst the semantic similarity of A and B is obvious, there are no terms shared between the two. Therefore, approaches such as TF-IDF would be ineffective.

	chrysanthemum	flower	lady	selling	woman
chrysanthemum	1.00	0.50	0.09	0.06	0.10
flower		1.00	0.10	0.09	0.11
lady			1.00	0.07	0.50
selling				1.00	0.09
woman					1.00

Table 6.2: Example term similarity matrix

Following preprocessing of the raw fields, the original *title* strings produce the vectors A = [chrysanthemum, lady] and B = [flower, selling, woman]. The results of using the maximum synset similarity to generate the term similarity matrix produce the results shown in table 6.2. As the table shows, 'chrysanthemum' and 'flower' have a high similarity (0.5) as would be expected, the same applies to 'lady' and 'woman' (0.5). However, unrelated terms such as 'chrysanthemum' and 'lady' have much lower values (0.09). The outcome of combining these weights with the values in the term vectors is shown in table 6.3.

With the weighted term vectors calculated it is now possible to calculate the cosine similarity. If this was done using the original term vectors then the result would be 0.00. However, if the similarity of the weighted vectors is calculated then a result of 0.76 is achieved.

		chrysanthemum	flower	lady	selling	woman
Term vectors	A	1	0	1	0	0
	B	0	1	0	1	1
Sim matrix values for A	chry...	1.00	0.50	0.09	0.06	0.10
	lady	0.09	0.10	1.00	0.07	0.50
Sim matrix values for B	flower	0.50	1.00	0.10	0.09	0.11
	selling	0.06	0.09	0.07	1.00	0.09
	woman	0.10	0.11	0.50	0.09	1.00
Weighted vectors	A	1.09	0.60	1.09	0.13	0.60
	B	0.66	1.20	0.67	1.18	1.20

Table 6.3: Example of original and corresponding weighted term vectors.

6.3.2.4 Conclusion

Although the conditional synsets approach produced better results than the maximum synsets approach used, the conditional synsets approach came with an unacceptable computational cost as it was not possible to cached the synset similarity values it produced for reuse in other *title* comparisons. This meant that the synset similarity values had to be redone for every single *title* comparison. In comparison, using the maximum synsets approach, once a pair of synsets had been compared their similarity value could be stored and reused very quickly which meant that the similarity metric as a whole ran much faster. When compared to the results produced by TF-IDF the inclusion of even this limited form of semantic similarity between terms is shown to be vital. A more in-depth analysis of the *title* fields when comparing would certainly produce better similarity values , but whether this could be achieved in a scalable manner is unknown at this time. At the present time, the pseudo-semantic similarity values described above represent a significant improvement on no weighted term vectors whilst still allowing for large scale processing.

6.3.3 *Description* field metric

Despite significant effort, it was not possible to produce a similarity metric for the *description* field. The problem is the sheer level of variation in the field, not just between different collections, but within single collections. Just within the ERPS collection the *description* field ranges between 0 and 717 words (compared to between 0 and 50 for *title*). The combination of large variations in field length,

contents and style posed significant challenges for automated analysis. It was not possible to overcome these in the available time frame.

While it was possible to just re-use the *title* metric and run it against the *description* field, the scarcity of populated fields and the sheer variety of information and lengths meant that resulting values rarely showed any resemblance to the perceived similarity of the overall record. At best, reusing the *title* metric to generate *description* similarity values represented a waste of processing time, at worst the near random nature of the similarity values generated could actively impede finding matches between the overall records.

The *description* field does often contain valuable additional information including details of technical processes, dates and locations. In the future it may be possible to extract information from this field in order to fill other missing elements of a record. However, at the present time the *description* fields are only used as one of the search fields in the initial blocking of the overall search space and pairwise similarities are not calculated for them.

6.3.4 *Person* field metric

Whilst name comparison is a common problem and there are, therefore, a large number of established solutions. However, GLAM collection records present a number of unusual challenges which are not typically encountered elsewhere or are at least not encountered together. These include...

1. Name order - The individual elements of the name can be stored in any order. Most name comparison systems assume that the individual elements of the name are stored in a known or at least the same order (i.e. 'first-name surname' or 'surname, first-name initial' etc). When comparing the *person* fields from GLAM records the format that the names are stored in changes, not just between different collections, but often within the same collection.

2. Short forms - E.g. initials instead of full names. Comparison of names which include short forms would not usually be an issue, the problem arises because of the need to be able to compare across forms (i.e. long vs short).

3. Additional information - Most commonly the inclusion of the birth and/or death years of the person in question. Whilst this can be valuable information, it does not belong in the *person* field.

6.3.4.1 Pre-processing

As with the *title* metric the first stage is to tokenise the incoming strings so as to produce a term vector of all the words in the supplied *person* fields. As with the *title* metric, low value words are removed Unlike *title* however, dates and other numbers are also removed. Filtering out titles such as 'Mr', 'Mrs' etc. was considered, however the conclusion was that these could act as a potentially valuable gender check on the name.

6.3.4.2 Similarity metric

Once the raw fields have been transformed into term vectors, the first stage is to produce an $|A| \cdot |B|$ similarity matrix of the terms. A and B in this case are the two term vectors and the similarity values are just the Jaro-Winkler[237] values for each pairwise comparison. On average there are only 3.15 words per field[38], so complete matrix generation is fast. An example matrix is shown in table 6.4. It is important for later stages that A be the smaller of the two vectors in terms of the number of elements. If both vectors have the same number of elements then it does not matter which one is A or B.

The next stage is to find the best match for each element of A to one of B with the added restriction that the matches are exclusive and so each element of B can only match a single one of A. The aim is to find the best overall match between the elements. Performing an exhaustive search of all possible combinations is too time consuming, even with the small size of the vectors mentioned earlier, since there will be $|A| \cdot |B|$ possible combinations. However by checking the most promising combinations first, the search time can be massively reduced, often to the point that only a single combination needs to be checked. This is achieved by ordering the $jarow(A, B)$ values for each A element (see table 6.5). At this point, a match for each element of A to one in B has been found.

The final step is to take the Jaro-Winkler[237] values for the pairs that were just selected and scale them according to the combined length of both elements in the pair. The scaling factor is calculated as the 1 divided by the total length of all elements in all of the best matching pairs. This scaling is important as it means that initials and shorter names are given less importance than full/longer names. For example, without the scaling a match between two elements both with the

[38] Analysis of 875,267 LoC records, 452,834 with non-null *person* fields produced a mean average of 3.15 words per field. Maximum length was 13 words.

value 'b' would be considered just as important as a match between two elements with the value of 'benjamin'.

A pseudo-code implementation of this approach can be found in algorithm 3 on page 208.

6.3.4.3 Worked example

In this example, the values of the *person* fields are "johnston, frances benjamin, 1864-1952"[39] and "miss frances b. johnston"[40]. These are real *person* fields taken from the LoC and ERPS collections respectively. Tokenisation and filtering produce the two vectors shown below.

- A = ['benjamin', 'frances', 'johnston']
- B = ['b', 'frances', 'johnston', 'miss']

	benjamin	frances	johnston
b	0.71	0.00	0.00
frances	0.35	1.00	0.51
johnston	0.47	0.51	1.00
miss	0.00	0.00	0.00

Table 6.4: Jaro-Winkler similarity matrix.

	benjamin		*frances*		*johnston*
b	0.71	frances	1.00	johnston	1.00
johnston	0.47	johnston	0.51	frances	0.51
frances	0.35	miss	0.00	miss	0.00
miss	0.00	b	0.00	b	0.00

Table 6.5: Ordered Jaro-Winkler similarity matrix.

As mentioned previously $|A| \leq |B|$ must be true. The Jaro-Winkler[237] similarity matrix for the vectors is shown in table 6.4. It would be possible to perform a comprehensive search of this matrix in order to find the best possible combination

[39] loc00651273
[40] erps17654

of matches but this would be inefficient. Table 6.5 shows the result of ordering the similarity matrix within each element of A, it clearly demonstrates that the best match for 'benjamin' is 'b', 'frances' matches 'frances' etc. Since there is no element of B that is the best match for multiple elements of A, no further combinations need to be checked. Even if there were a match collision, the ordering of the similarity values makes it trivial to search through only the most promising combinations in order to find one that is lacking any collisions. This demonstrates that ordering the similarity matrices massively reduces the number of combinations which need to be tested and so dramatically reduces processing time.

With the A, B matches identified, the match weighting can begin. Table 6.6 shows the initial match values, the average weight of the matching elements, the weight values which are applied to the initial match values, the result of combining the match and weight values and finally the overall similarity value for the two vectors. Using this approach the two example *person* fields produce a similarity value of 0.93.

	benjamin b	frances frances	johnston johnston
Jaro-Winkler	0.71	1.00	1.00
Length	4.5	6	8
Weight	0.23	0.36	0.41
Combined	0.16	0.36	0.41
Result			0.93

Table 6.6: *Person* field similarity metric result.

The effectiveness of the approach decreases as the number of elements in the *person* fields increases, both in terms of the accuracy of the similarity values produces and in terms of increasing processing time. This loss of performance is due to the increasing numbers of elements offering a greater chance that pairs of elements being compared will have high similarity values by random chance. Such random matches causes issues in that two elements that are not actually the same will match and thus skew the final similarity value. They also increase the chance of an element from one vector matching multiple element matches from the other. Element match collisions require that additional combinations of elements be tested which increases the processing time.

6.3.4.4 Conclusion

Although the *person* process can encounter difficulties with longer values, its performance on short to medium length fields is more than acceptable. Its ability to compare despite differing name orders and between full names and initials make it very well suited to the *person* fields of GLAM collections.

6.3.5 *Process* field metric

Of all the fields for which a similarity metric was successfully produced, *process* was the most difficult.

The majority of historical photographs required multiple chemical processes to be carried out in order to produce what most lay individuals would describe as a photograph (a positive print of an image). Typically one set of processing was done to create a photographic negative and then another lot of processing was done to create a positive print from that negative. However, for many in the GLAM communities, "photograph" can refer to photographic negatives, reproduction prints, enlargements, positive prints and the results of historical processes with no good modern day analogue (e.g. daguerreotypes), with different community individuals disagreeing on which of these qualify as photographs. Few *process* fields list all of the processes that were used to create the physical item stored in the collection. Therefore even when two collection items were created using the same processes, what is actually listed in the record metadata may be completely different.

The other significant issue encountered was that the accuracy of the information stored in *process* is notoriously poor. This is due to high levels of miss-identification of the photographic processes used, in part because multiple processes can all produce very similar outputs and the information needed to distinguish between them is not captured in digital copies of the images[41][180].

Interestingly, whilst the problem of process misidentification is often referred to in the literature, no evidence of any research, which has been conducted into solving the issue from a search standpoint, could be found. The only investigations which remotely touched on the subject at the time of this research appeared to be research into developing expanding the list of known photographic processes[207] and into methods of identifying the photographic processes used for an individual

[41]For example the texture of a photograph is often used to identify Albumen prints.

photograph[206]. Both of these require access to the physical photographs. It is thought that no-one has analysed the rates of misidentification within collections. At the present time, knowledge of misidentification consists of the intuition of photo-historians based on their previous experience with collections.

Given the absence of any existing statistics on misidentification rates, a different approach was required. The inspiration and basis for the approach that was developed was the book "Care and identification of 19th-century photographic prints"[180], which provides a fold out flowchart describing how to identify various photographic processes. The steps described in that flowchart were unfortunately unsuitable for use in the *process* metric. The flowchart only covers a portion of the processes listed in the ≈1.7 million records collected and relies heavily on photographic characteristics which either can't, or are unlikely to be captured in digital copies. Characteristics such as whether the photograph is matte or glossy and the texture of the surface of the image. Although this layout was eventually discard completely, the layout of the photographic processes within the flowchart was copied to form the first attempt at a hierarchical representation of the processes and their relationships to one another and provided the initial inspiration for representing process similarity in terms of hierarchical distance.

An unexpected problem was the number of non-photographic processes listed in the records collected and how these should be handled. In some cases the problem was simply that a record had been miss-classified, records for paintings, cups, badges and valentine cards[42]. While these records did have photographs of the artefacts, the records were classified as being for a photographic artefact, i.e. the underlying artefact was a photograph. One complication is that it is not possible to identify between misclassified records and records which are correctly classified but where the process field describes the subject of the photograph rather that the process behind the photograph. Since anything can be the subject of a photograph and GLAMs can have almost anything in their collections, in these cases the *process* field could say anything. It was simply not possible to include every type of object and/or material in the world. Less clear were records related to pieces of photographic paraphernalia, camera, lenses and cases rather than actual photographs[43].

[42]E.g. loc2010645779, `http://www.loc.gov/pictures/item/2010645779/`. In this case the *process* field contained "1 item : lace, color."

[43]E.g. vaO123776, `http://collections.vam.ac.uk/item/O123776/`. In this case the *process* field contained "Walnut and cuban mahogany with ebony and boxwood inlay". This record has since been re-classified correctly as furniture.

Finally there were processes such as halftone printing and lithography, these are printing and not photographic techniques but which are strongly connected to photography. These processes were not specifically addressed in the process hierarchy but may still match against it. For example "dry plate camera" would match against "dry plate", one of the keyword sets for "Gelatin silver" process in the hierarchy[44].

These issues were only resolved following significant discussions with experts in the field of photographic history, primarily Professor Roger Taylor[45] and Dr Kelly Wilder[46]. In consultation with Professor Tayor and Dr Wilder, the concepts and process hierarchy which form the *process* metric were gradually developed. The final method is built on six main concepts.

1. Different names for the same process are given a similarity of 1.0 to each other. A single process may be listed under different names in different records due to the preference of the record's author or institution. Commercial names for a single underlying process are not uncommon[47].

2. Two processes of the same overall type should be considered more similar than two processes of different types[48]. For example, Albumens and Collotypes are positive images on paper, therefore they should have a greater similarity than between Albumen and Collodions since the latter is a negative print on glass.

3. The similarities between the process types are not all equal. For example, paper positives and paper negatives have a greater similarity than paper positives and direct positives.

4. A single process can belong to more than one process type. For example, Calotype prints can be both paper positives and paper negatives.

5. When possible, a *process* field should be matched to a process type even if it is not possible to match it to a specific process. For example, a *process* containing

[44]See section D on page 210.

[45]One of the leading experts in the study of 19$^{\text{th}}$ century photographic processes[158] and creator of the PEiB collection[221] utilised in this research.

[46]Reader in Photographic History at De Montfort University.

[47]E.g. Collocolour/Mezzograph (Collotype), Real Photo (Gelatin Silver Print) and Sepiatype (Vandyke Print)[16, 208].

[48]Process type is used to indicate a collection of different processes (e.g. Albumen, Kallitype, Collotype) which are grouped according to whether they produce positive or negative images and the material on which the image is produced (i.e. paper, glass).

"positive print on paper" clearly belongs to the paper positive process type, but identifying a particular process within that type is unachievable.

6. It is not possible to identify the actual processes which were used without access to the physical records and, even then, it would not be possible to identify the processes in an automated manner. Therefore no allowance is made for the fact that a pair of records might have had one or both processes misidentified as the same process. Such situations would erroneously produce a similarity of 1.0.

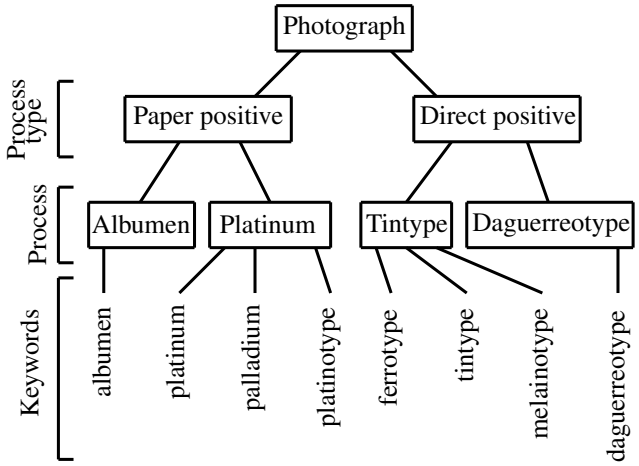


Fig. 6.6: Subset of the processes hierarchy used by the *process* metric.

6.3.5.1 Pre-processing

When pre-processing the *process* fields the aims was to identify occurrences of known photographic processes in the fields and to convert all instances of a particular process to a single standard representation. This was achieved by matching the contents of the *process* fields against a list of keywords. Each photographic process had one or more sets of associated keywords. For example, the Tintype process is associated with the keywords "tintype", "ferrotype" and "melainotype". The tokenised *process* fields were compared to the keyword sets for all processes. Matches between the tokenised words and the keywords are performed using Jaro-Winkler[237]. When a keyword set contained more than one word, the overall set match was calculated as the mean of all keywords in the set.

The highest matching keyword set with a match value of ≥ 0.85 is selected as the process for each *process* field. This threshold value allows minor spelling

Process	Keywords	Match scores vs. 'platinotypes'
Platinum print	platinotype	0.98
	palladiotype	0.77
Tintype	ferrotype	0.59
Salted paper print	salted, paper	$\frac{1}{2}(0.47 + 0.57) = 0.52$
	silver, chloride	$\frac{1}{2}(0.33 + 0.53) = 0.43$

Table 6.7: Example comparison of *process* field containing 'platinotypes' against a subsection of process keyword sets.

	Paper positive	Paper negative	Glass negative	Direct positive
Paper positive	0.75	0.50	0.50	0.25
Paper negative		0.75	0.25	0.25
Glass negative			0.75	0.25
Direct positive				0.75

Table 6.8: Process types similarity values.

mistakes and the plural forms of words to match the listed keywords. An example of matching a field containing 'platinotypes' to a subset of the process hierarchy is shown in table 6.7. In this example, the *process* fields will match correctly to the Platinum print process.

6.3.5.2 Similarity metric

Once the actual processes listed in each *process* field have been identified the next step is to identify the similarities between them. For this, inspiration was taken from WordNet and the *title* metric. Instead of measuring word similarity as a function of path distance across the synset hierarchy, it was measured as a modified path distance across the process hierarchy. Consequently, those processes which are closer together within the hierarchy have a greater similarity than those distantly placed within the hierarchy.

For the path distance approach, the shortest possible path from process A to B is found, measured as the number of edges that the path traverses. The maximum of the edge weights along that path is then used as the final similarity between A

and B.

6.3.5.3 Conclusion

The main concern with this metric is that the similarity values for the high level photographic types (i.e. paper positive, glass negative etc.) were arrived at manually by a combination of trial-and-error and consultation with Professor Taylor and Dr Wilder. They were not the result of a statistical analysis of real world process misidentification rates. The issue is that, lacking a fuller investigation, the values that were used may not have accurately described the relative rates of misidentification. A full investigation in this case would mean an in depth study of, at the very least, thousands of photographic items from across as broad a set of photographic collection as possible. Each records would need to be examined in detail, to ensure that the photographic process listed in the item's metadata are correct. The relative proportions of the various processes recorded and rates of photographic process miss-identification for each process calculated.

In addition, the process hierarchy contains only the small subsection of known processes, those which are actually present within the ≈1.7 million records used in this research and is, therefore, not an exhaustive list of all photographic processes. Use of this metric in a wider context would require further expansion of the processes list and associated keywords and ideally updates to the process similarity values based on statistical analysis of the real rates of process misidentification within various collections. However, for this research a simple hierarchy of fifty sets of keywords linked to twenty three different process types and four overall process groups was sufficient to cover the processes present in the ≈1.7 million records. The manually developed similarity weights, whilst not ideal, appear effective. The lack of any known previous attempts to model process misidentification in collections means that this metric must be considered an improvement on the current state of searching (pure keyword searching).

The full set of processes, relationships and keywords can be found in section D on page 210.

6.3.6 *Date* field metric

There are two significant factors which must be considered when trying to compare *date* fields.

1. The time difference between the two *dates*. The earliest date listed in the records used in this research is "5000 BC - 4500 BC", at the other end of the scale records which claim to be from "22/12/2559"[49] exist. These are obviously errors in the collection information, or in the case of the BC dates, photographs of collection items from that time (pots, coins etc) which have been misidentified as photographic collection items. Nevertheless these extreme dates do demonstrate that there is a significant time span over which to compare dates. Taking as the starting point the date of the first images recorded permanently using chemicals (≈1790[50]) and 2013 as the end point, that still leaves a period of more than two centuries within which the records could fall.

2. The time span covered by the two fields. For example, ≈90,000 *date* fields from the LoC collection had an average span of ≈4.7 years

If the date ranges described in two fields are the same then clearly those are more similar than two that describe different ranges. Unless, the ranges in the later pair describe a narrower range of dates. For example, '1900 to 2000' vs. '1900 to 2000' should be considered less similar than '1900 to 1910' vs. '1905 to 1915'.

6.3.6.1 Pre-processing

The fields are being collected from a number of institutions, each of which can use a different date format, multiple date formats or have no consistent date format. Therefore the decision was made not to assume the format for any *date* field regardless of originating collection. The formats of each field would instead be deduced from scratch. This was achieved through a combination of rules, regular expressions (regexes) and the python dateutil library[154].

The major challenge was dealing with date ranges that were specified in conversational styles. For example, "the 19th century" or "1890s". When at all possible the rules used were designed to convert these into '1800 to 1899' [51] and '1890 to 1899' respectively.

Only the years of the extracted dates were used in the *date* metric. More precise information was sometimes available but only for a very small minority of the records

[49] nz1377614 amongst others

[50] Typically attributed to Thomas Wedgwood, these earliest images could only record an object silhouette (photograms) via direct contact with the light sensitive surface.

[51] Centuries technically run from 1801 to 1900 etc, however the popular misunderstanding is that they run 1800 to 1899 etc. The rules operate according to the misunderstanding.

(0.004%[52]). Where precise date information was available (i.e. a specific day or month), this was not considered, only the year information was used. There were several reasons for this, firstly the accuracy of the information. A suspiciously high number of photographs are apparently taken on January the 1st of each year, the more likely explanation is that software packages used to create the records default to January the 1st when day and month information is unknown. Secondly, it is often impossible to identify between *date* fields using DD/MM/YYYY formats and those using the American MM/DD/YYYY format. That is to say that if the *date* field contains 03/05/2001 it is impossible to know if that is the 3rd of May or the 5th of April. Thirdly, the majority of *date* fields don't contain day level information. Working on the assumption that 1st of January dates are not valid, only 29% of *date* fields contain day level information[53] Fourth, the difference this made to the results from the *date* metric was at most 0.02 or 2%. Given the insignificant improvements in accuracy, that most fields did not have this additional information and that it could not be trusted even when it was present, it was simpler and easier to just ignore the day and month data.

6.3.6.2 Similarity metric

The *date* similarity metric is the simplest of the individual field metrics to calculate. It takes three inputs; two date ranges (A and B) and a span weight (y). Each date range has a starting year (A_s) and an ending year (A_e), these are used to calculate the span of the range (A_p, see equation 6.2).

The *date* similarity is constructed from three sub dissimilarities...

1. p - The mean average span of the two date ranges.
2. s - The difference between when the two date ranges start.
3. e - The difference between when the two date ranges end.

[52] 2,094 out of 574,631 LoC records.

[53] Based on an analysis of 4,253,685 *date* fields, date information was contained/could be identified in 1,336,666 and day level information was found in 373,856.

$$A_p = A_e - A_s + 1 \tag{6.1}$$

$$p = \min\left(1, \max\left(0, \frac{\frac{1}{2}(A_p + B_p)}{\Phi}\right)\right) \tag{6.2}$$

$$s = \min\left(1, \max\left(0, \frac{|A_s - B_s|}{\Phi}\right)\right) \tag{6.3}$$

$$e = \min\left(1, \max\left(0, \frac{|A_e - B_e|}{\Phi}\right)\right) \tag{6.4}$$

$$\text{datesim}() = 1 - \frac{p + s + e}{3} \tag{6.5}$$

All three of these components are scaled using the Φ value supplied and restricted to a range of [0 1]. The Φ value controls how forgiving the metric is of the differing time spans. For the purpose of this research $\Phi = 50$, a value arrived at by trial and error. This means that if, for example, the two date ranges had an average time span of 50 years, the p dissimilarity would be 1. A similarity of 0 for date ranges over fifty years may seem extreme and many GLAM records do have longer time spans listed (e.g. "19$^{\text{th}}$ century") but photography as a viable and widely used technology has existed for less than two hundred years. If the date ranges cannot place a photograph's origins more precisely than a quarter of the entire history of the technology, then the usefulness of that piece of information for co-reference identification is extremely limited.

6.3.6.3 Worked example

In this example, the date ranges used are:

$$A_s = 1888, A_e = 1910$$

$$B_s = 1874, B_e = 1897$$

See figure 6.7 for a visual representation of the example date ranges and resulting start/end gaps. Using these values therefore, A_p and B_p equal 22 and 23 respectively to produce an overall p value of 0.45. The start gap of the two dates is 14 years, as such $s = (|1888 - 1874|)/50 = 14/50 = 0.28$. Similarly the end gap is 13 years, so $e = (|1910 - 1897|)/50 = 13/50 = 0.26$. Combining these produces an overall dissimilarity value of $\frac{1}{3}(0.45 + 0.28 + 0.26) = 0.33$. The overall similarity value is therefore $1.0 - 0.33 = 0.67$.

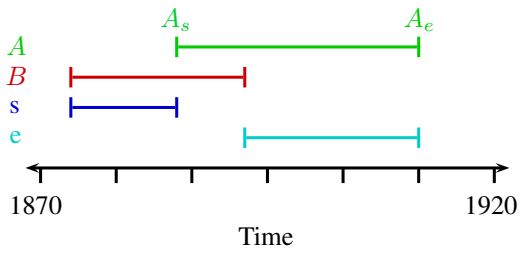


Fig. 6.7: Example date ranges and gaps.

6.3.6.4 Conclusion

Whilst the approach is able to compare successfully date ranges, and is able to understand a significant proportion of the date formats which are used in GLAM collections, it is not able to understand all formats. For example "c186-?"[54], "25 feb ?"[55], "pre world war two"[56] and "late 20 century"[57].

The formats which can not be processed are fairly unusual and, as data entry standards begin to be enforced and employed by GLAM institutions[106] these formats can be expected to disappear. It should be noted that in several cases, the copies of these records held by the external collections have already been corrected/updated and the errors only persist in the local copies of them.

6.4 Overall record similarity

As discussed previously (section 3 on page 34), the lack of a GUID for photographs across GLAM collections and the imprecise nature of the individual fields in the records, means that no single field can be used to compare successfully the records even if the added complication of uncertain field matches were not present. Therefore in order to arrive successfully at an accurate overall record similarity, multiple field comparisons need to be combined.

At this stage in the process, each record pair has a set of four similarity values which represent the similarities between each of the individual fields. The earliest attempts at an overall record similarity value were based around field similarity averages, maxima, minima and combined sums. It was immediately apparent that these

[54] loc2003670698.
[55] nz23264416.
[56] nz20364130.
[57] nz29850871.

approaches were not very successful and so they were quickly discarded. However, they were used for the early VAT images (see sections 3.4.4.1 on page 55 and 6.5.1 on page 123).

It was clear from an early stage that certain fields had a greater importance from an overall record match than others. PRL and ANN all make clear allowances for the fact that certain inputs (in this case field similarities) can be significantly more valuable than others and so it was not unexpected when differences in field importance became apparent. The GLAM community questionnaire also highlighted that certain fields saw, not just far greater use than others, but also that the level of importance placed on the fields was similarly varied. Initially an approach similar to PRL using hand coded field weights was attempted in the hope that as there were only a very small number of fields, it would be possible to arrive at satisfactory values through trial and error. However it quickly became clear from the results produced that this approach was fundamentally unsuitable.

As a naive Bayes Classifier, PRL did not allow conditional dependence to be modelled and instead assumed that each field is conditionally independent. This was already known to be incorrect for GLAM records. For instance, the *person*, *process* and *date* fields were all expected to affect each other. Relatively weak relationships such as this does not, however, exclude the use of naive Bayes Classifiers.

In practice what occurred with PRL was that high similarities the less important fields[58] had a significant effect on the overall record similarities. Modifying the field weightings to apply sufficient weight to the important fields to prevent this happening meant that the less, but still potentially useful, fields were unable to affect the overall similarity values in cases when the more important fields[59] were not producing good matches. The record ordering which resulted was unsatisfactory and so the PRL based approach was abandoned.

6.4.1 Fuzzy Inference System (FIS)

It was clear from the early approaches that the individual fields were dependent and that treating the field similarity values as independent variables was not producing the desired results. The problem was not suitable for naive Bayes classifiers. Since supervised learning approaches were not an option, a rule based approaches utilising the similarity values of the individual fields as the inputs was decided

[58] *Process* and *date*.
[59] *Title* and *person*.

upon. As discussed earlier[60], rules based approach often function poorly when faced with imprecise and uncertain information. In order to mitigate the issue of the field similarity value uncertainty and to simplify the rule creation, a fuzzy logic approach is used. The overall record similarity is, therefore, the output of a FIS.

The similarity values from the individual fields are all fuzzified using the same two fuzzy sets consisting of a 'good' and 'bad' set. The output consists of three sets describing 'good', 'ok' and 'bad' overall matches. The sets used can be seen in figure 6.8.

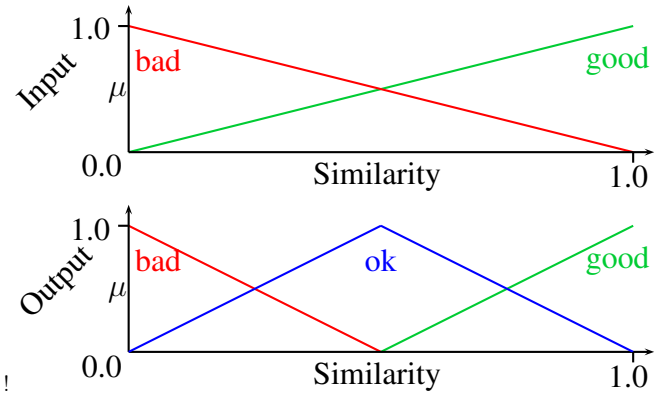


Fig. 6.8: Fuzzy sets used for overall record similarity.

The full set of fuzzy rules used is as follows;

IF *title* is good **AND** *person* is good **THEN** match is good.

IF *title* is good **AND** (*date* is good **OR** *process* is good) **THEN** match is ok.

IF *person* is good **AND** *title* is bad **THEN** match is ok.

IF *title* is bad **AND** *person* is bad **THEN** match is bad.

These rules place a greater significance on the *title* and *person* similarity values. The *process* and *date* similarities are used but only to reinforce already acceptable matches based on the *title* and/or *person* similarities. These rules were arrived at

[60]See section 3.1.

through a combination of trial and error and analysis of the questionnaire responses received from members of the GLAM community.

As it was not possible to develop a satisfactory *description* metric (see section 6.3.3), this field was not used. However as the results of the online questionnaire show (see section 1.2), the relative importances placed on the fields in the fuzzy rules is representative of the views of the GLAM community members questioned.

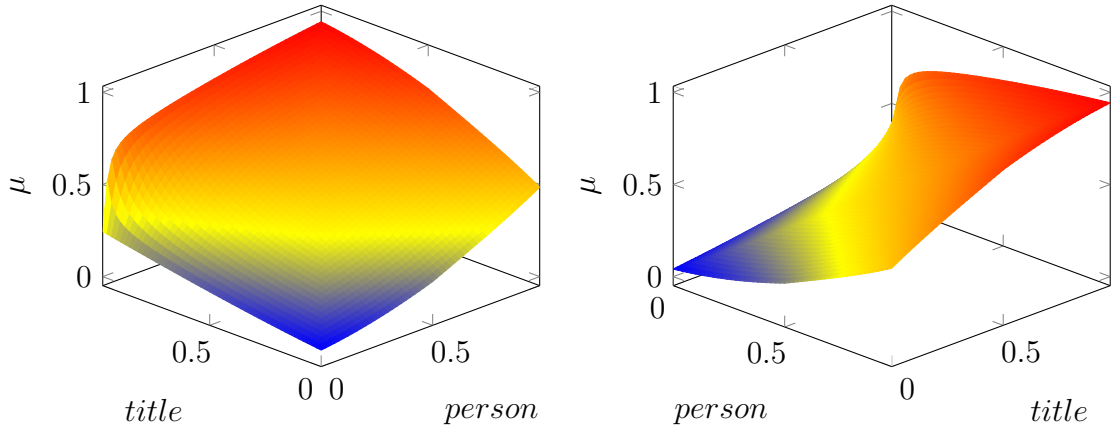


Fig. 6.9: FIS output surface, inputs *title* and *person*, *process* = *date* = 0.0.

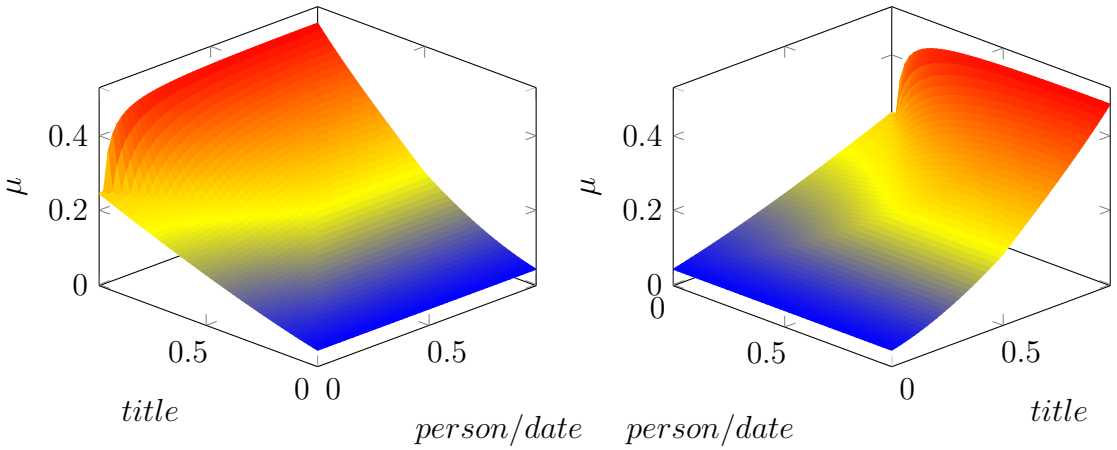


Fig. 6.10: FIS output surface, inputs *title* and *process*/*date*, *person* = 0.0. Note that the lack of a *person* similarity value produces significantly lower results than those seen in figure 6.9.

6.5 Record ordering

With the overall record similarity matrix generated, a comprehensive list of the estimated similarity of each record to every other record is available. However, even

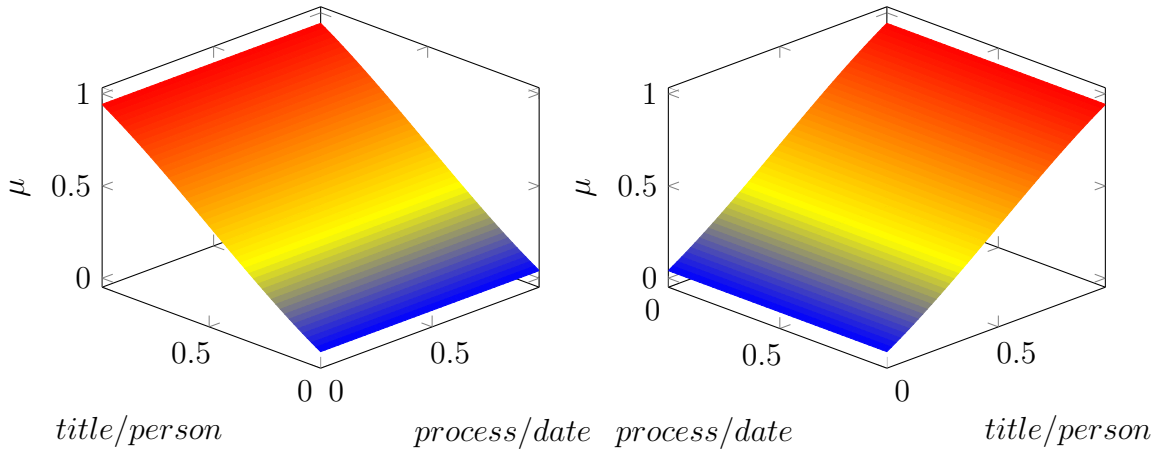


Fig. 6.11: FIS output surface, inputs *title/person* and *process/date*. Note that, in line with their perceived importance, change in *title/person* have a far more significant effect on the output than changes in *process/date*.

a small similarity matrix contains far too much information to be examined/understood manually. It is necessary to identify and extract the most promising records in the matrix and then present those records to the users for further consideration.

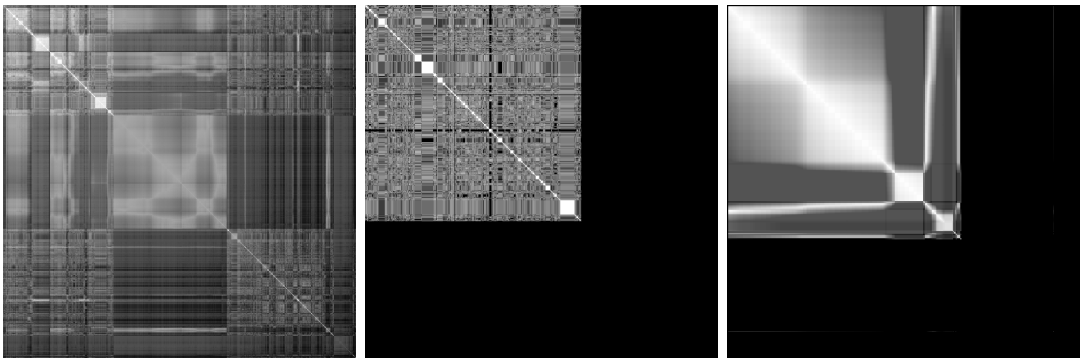
Several different approaches were explored in order to arrive at a satisfactory solution. The initial experiments were focused on clustering the record similarity matrix, but this did not achieve the results required. In the end, a constrained depth first search algorithm was resorted to in order to generate a dendrogram of the records. The failed clustering attempts are discussed in section 6.5.1 including the reasoning behind this approach and ultimately the reasons why it failed. In section 6.5.2 on the following page the constrained depth first approach and the results it produces are described.

6.5.1 Clustering

In order to analyse the suitability of the data for clustering, a series of VAT images were produced[61]. The VAT images produced showed clear structure in the similarity values, at the level of both individual fields and overall records, and this led to a belief that clustering would be an effective approach. Unfortunately, this proved to be incorrect and clustering ultimately proved to be unsuitable. Whilst clusters were identified and the contained records did resemble each other, the similarities between

[61] See section 3.4.4.1.

records due to the semantic meanings of their *title* text etc. were overpowered by the similarities between records from the same collections. If two random records from different collections and two from the same collection were compared, on average the records from the same collection would have a greater similarity than the separate collection ones. This is due to the same field formats and terminology appearing multiple times in the same collections. The issue was that this 'background similarity' between records from the same collections was having a noticeable effect on the clusters being produced, distorting the clusters and prevented the identification of co-referent records across different collections.



(a) Overall record. (b) *person*. (c) *date*.

Fig. 6.12: VAT images of various erps17093 similarity matrices. Note the bright white squares in each of the images, these show that there are a potentially (but not definitely) a number of strong clusters in the similarity matrices.

6.5.2 Constrained search

The failure with clustering necessitated a fundamentally different treatment of the records. With clustering, the seed record was treated as just another record in the search space to be clustered. The new approach applies a specific focus to the seed record. The overall aim is to find those records which are potentially co-referent with the seed record and, therefore, it makes sense to identify the records with the highest similarity to the seed record first.

The approach used to achieve this is similar to depth first searching and Minimum Spanning Trees (MSTs). A MST is a sub-graph that has no closed loops (is a tree), which connects to every single node within the full graph (is a spanning tree) and where the sum of the edge weights is less than or equal to every other possible sub-graph.

The approach starts with the seed record at the root node and adds as a child node the record with the highest similarity to the seed record. The search sub-set record with the highest similarity to either of those two nodes is then added as a child node of the record it has the highest similarity to. The process now iterates, with the record with the highest similarity to the records already in the graph being added as a child node. Eventually all of the records will have been added to the tree and the process ends. An example implementation of this process can be found in algorithm 4 on page 214.

The end result is that those records with the greatest similarity to the seed record appear in the highest layers of the spanning tree. Of course a similar effect could be achieved by simply selecting the n records with the highest similarity to the seed record. However, this approach does more than just select the records with the greatest similarity to the seed record, it also groups similar records together within the hierarchy. For example, records with the same/similar *person* fields will be grouped together which allows for easy exploration of all the records by a single photographer.

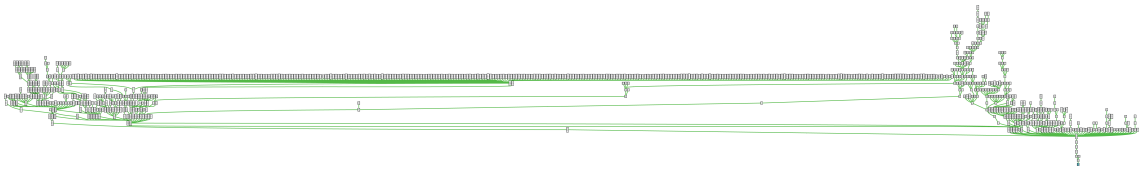


Fig. 6.13: Example result, full graph for ERPS record 17093 (rotated clockwise 90°). Note that this figure is not intended to show the dendrogram in any detail, just the overall size and shape. For a more detailed view of the top portion of the dendrogram, see figure 6.14 on the next page.

The graphs produced by the algorithm contain the same number of nodes as the records which were processed. As a result, visualising the graphs becomes progressively harder as the number being compared increases. Since the estimated similarities between the records and the seed record decreases in relation to the number of edges between the nodes as the root, distant nodes can be safely discarded if necessary. Figure 6.14 shows the effect of discarding all but the top 100 nodes of the graph shown in figure 6.13. The 'top' nodes in this case are defined as the first n records to be added to the graph.

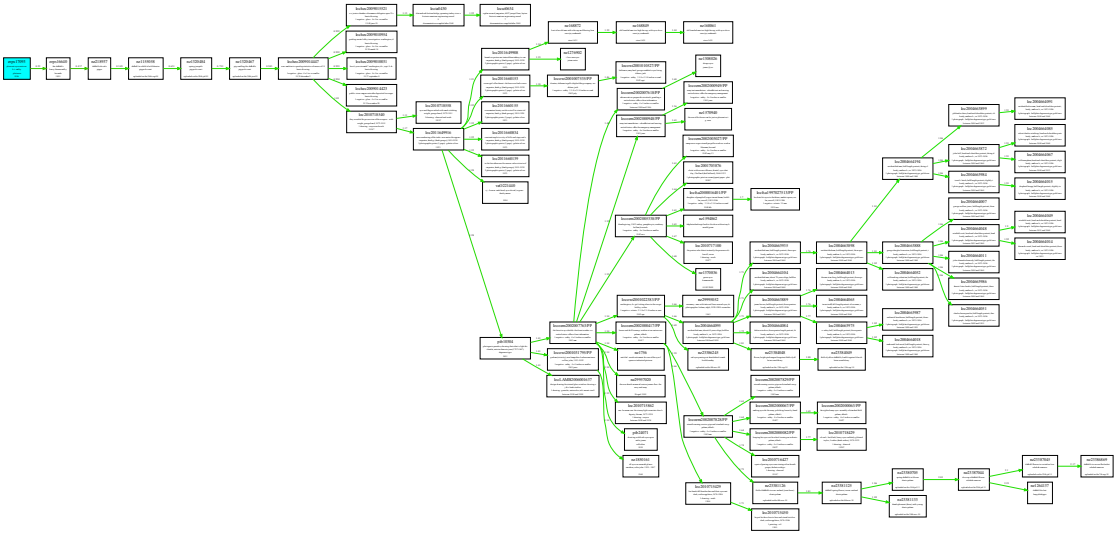


Fig. 6.14: Example result, top 100 results for ERPS record 17093. Note that this figure is not intended to show the records in detail, just a improved view of the hierarchical structure than can be see in figure 6.13 on the preceding page. For a figure which shows record details please see figure 6.15.

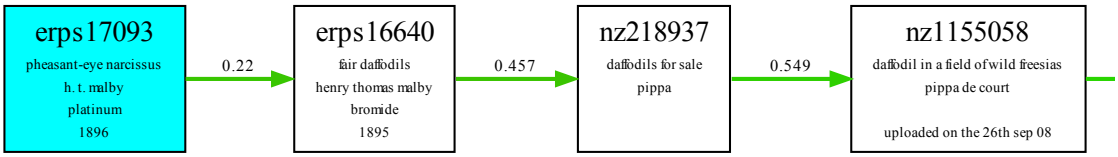


Fig. 6.15: Example result, top 4 results for ERPS record 17093.

6.6 Summary

In the end, there is simply not enough information available in the ERPS records (or in many of the external collections) to be able to state definitively when two records are co-referent. With further investigations and additional information (receipts, auction records etc.) it may be possible to build a convincing case that two records are referring to the same photograph but this information is either not available online, or would prove incredibly difficult to search with an automated system. However, the ordering of the records which the constrained search system produces, means that the most likely co-reference candidates are brought to the fore which in turn means that it is significantly easier to identify records of potential interest than would be the case for collections which return unordered results. Even for those collections that do perform simple record ordering based on term counts, the fuzzy logic similarity measure described above is significantly more effective at

identifying underlying record similarity. This means better record ordering using the fuzzy approach which will be demonstrated during the testing chapter[62].

Whilst it is possible to simply select the top n records/layers from the dendrogram or apply a static threshold and state that these records are believed to be co-referent, realistically any claim of co-reference needs to be backed up by a significant amount of detective work, ideally by trained photo-historians. While the approach cannot identify co-referent records on its own, it can successfully order the records based on the co-reference likelihood, significantly reduce the number of records which need to be examined by photo-historians and, therefore, reduce their workload. This means that the proposed approach could also be considered an information filtering or recommender system[182].

[62]See section 7.

7

Testing

The central question of this research was whether it was, or was not possible to locate the 'missing' images from the ERPS collection without resorting to a manual search and the excessive person-hour requirements that would entail[1]. The previous chapter describes one possible approach for locating these images. This new approach mimics in many ways the search techniques used by manual searchers. Since the approach proposed is not fully automatic, some manual involvement by photo-history researchers is still required in order to find the ERPS images. However the person hour requirements of the proposed approach should be lower than those of manual searching as the majority of the search process is automated, i.e. identifying the keywords, visiting the collections, collecting the records and identifying likely co-reference candidates[2]. The amount of time needed to locate potential co-reference matches within the collections is not, however, part of this testing.

The source code produced during this research was written so as to easy to change as possible and to log the results from every intermediary stage. The intention was to provide plenty of flexibility to experiment with different algorithms/approaches and to assist in identifying where mistakes and/or errors had occurred. As a consequence, processing efficiency was sacrificed and the current code is highly inefficient. The end result is that any measurements that were made of the amount of time needed for the proposed approach to identify potential matches is unlikely to have any resemblance to the time required for a fully developed, real world implementation. While the overall processing time of the proposed approach was kept in mind throughout this research[3], the code that combines the results of the individual similarity metrics, produces the overall record similarity and produces the final record match results was not written to be computationally efficient. Therefore, the amount of time

[1] See section 3.5 on page 57.

[2] See section 6 on page 87.

[3] I.e. in the development of the *title* metric, see sections 6.3.2 on page 100 and 7.2 on page 151.

needed to search using each approach was not recorded, as any comparisons at this time would be inaccurate and meaningless.

If at the end of testing the proposed approach has been shown to be capable of identifying high quality co-referent record matches, then future work would involve rewriting the source code into a single, computationally efficient process. This would not, however, have been a sensible use of resources until the proposed approach has proven the quality of its record matching.

The success of the proposed approach with regards to this testing is therefore based on one factor, that the proposed approach is actually able to find the missing ERPS images. Any time savings would be irrelevant if it was not possible to locate the necessary photographic matches in the first place. The performance of the proposed approach with regards to its record finding abilities when searching and what that means for its usefulness can be summed up by the following four possibilities:

1. The proposed approach is totally unable to find and/or bring to the attention of the photo-historians searching, any photographic records which match any of the ERPS images. However a manual search found suitable matches. Under these circumstances, the proposed approach has failed to achieve its main function, locating relevant photographic records and would either need to be modified until it could or it would need to be discarded.

2. The proposed approach is able to find and bring attention to some photographic records, but a manual search was able to locate significantly more or find ones of a higher quality. If this had occurred, then the usefulness of the proposed approach would be dependent on a combination of how much more effective manual searching was and how much of a time saving the proposed approach offers. If manual searching offers matches of significantly greater quality with only minor man-hour increases compared to the proposed approach, then the proposed approach would not constitute an effective substitute. However if manual searching corresponds to a minor improvement in record matching for a significant increase in man-hours required then an argument could be made for the proposed approach.

3. Both the proposed approach and manually searching are able to find potential matches but neither approach produces significantly more, or significantly better matches. In this circumstance, the man-hour savings of the proposed approach would mean that the proposed approach offers a clear benefit compared to manual searching.

4. The proposed approach is able to identify significantly more and/or better matches than manual searching. In this situation, there are indisputable benefits to the use of the proposed approach with regards to recall and/or quality. Improvements in man-hour requirements would only enhance the suitability of the proposed approach.

As the four possible situations show, the effectiveness of the new approach is dependent on its ability to find and present relevant results and the quality of the results that it finds.

The intention of the testing that was conducted as part of this research was to determine if the proposed approach is able to produce results that are equivalent to, or represent an improvement on, those of manual searching. As long as a minimum level of result quality is achieved[4], then the proposed approach can be considered a success on the basis of the predicted time savings due to the automation of most of the search tasks[5] and the significant time requirements for manual searching[6]. Further refinements and improvements conducted as part of future work/research.

7.1 Result quality

Testing the quality of the results produced by the new approach was a challenge. The typical and easiest way to determine the performance of a new search or co-reference identification approach is to use a pre-existing gold standard dataset to measure directly the recall and precision rates that are achieved[7]. A gold standard dataset represents the ideal results. No approach can achieve results of the same quality in the real world for reasons of cost, time, equipment etc. Although it is generally impossible to match the results of gold standard datasets, any other approach can be easily measured against it and multiple approaches can, therefore, be easily contrasted. Unfortunately this approach was not an option for this research for two reasons.

Firstly and as previously discussed[8], there was no suitable pre-existing labelled dataset available for training or testing co-reference identification systems for GLAM records and creating one from scratch was not a realistic option given the time and

[4]Possibility 2 on the preceding page.
[5]See section 6 on page 87
[6]See section 3.5 on page 57.
[7]See section 2 for explanations of recall and precision.
[8]See section 3.5.

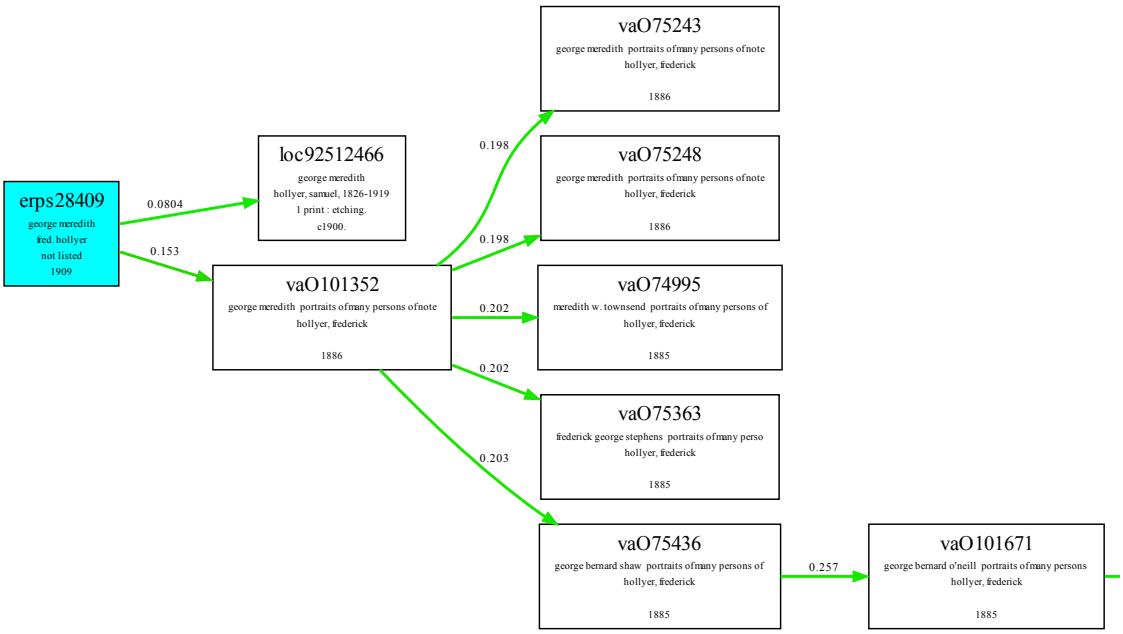


Fig. 7.1: Top records of the erps28409 dendrogram.

resources available for this research.

Secondly, measuring the precision would require that it could be definitively stated if two records were, or were not, co-referent. However there was no absolute, ground truth measure of co-reference in this case. If both records are referring to the exact same physical photograph then that was clearly a match. This was, however, an incredibly narrow focus and for most searches the criteria for what constitutes a match could be significantly broader. What constitutes a good match between photographs varies depending on the individual looking at them. For example, an original photograph and an enlargement of the same image may or may not constitute matches depending on whether the searcher was interested in the photographic content or the photographic processes.

To give a real example, table 7.1 shows the images associated with three records found in the 1.7 million GLAM records that were collected[9], erps28409, loc92512466 and vaO75248. Figure 7.1 shows a subset of the dendrogram from which these records were selected. Based on the metadata for the records it appears that vaO75248 was taken by Frederick Hollyer in 1886, then later erps28409 was created for exhibition in 1909 by enlarging of a portion of vaO75248. An etching was also produced by Frederick's brother Samuel in 1900 which resulted in loc92512466. Whilst all of these records are clearly closely related, they are not the exact same

[9]See section 6.2.1 on page 94.

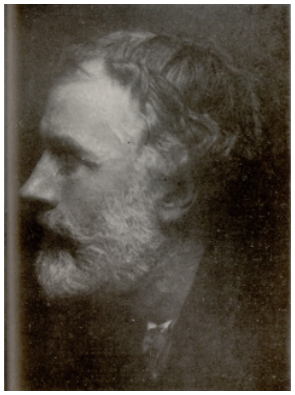
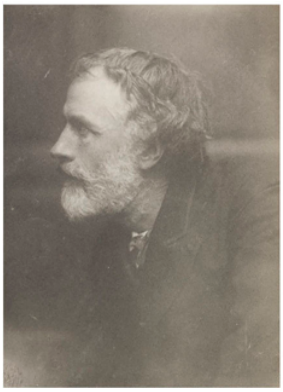
Id	erps28409	loc92512466	vaO75248
Title	George Meredith	George Meredith	George Meredith; Portraits of many persons of note photographed by Frederick Hollyer
Person	Fred. Hollyer	Hollyer, Samuel, 1826-1919	Hollyer, Frederick
Process	[Not Listed]	1 print : etching.	
Date	1909	c1900.	1886
Image			
Found by	N/A	Both	Both
Attribution	Copyright ©2008 De Montfort University. Database right De Montfort University (Maker). All rights reserved.	Courtesy of the Library of Congress, LC-USZ62-105804.	©Victoria and Albert Museum, London.

Table 7.1: Co-reference candidates for erps28409.

physical photograph. Finding such close connections for erps28409 may, however, still be a valuable result. An investigation into George Meredith would likely consider all of the images in table table 7.1 very strong matches. However an investigation based on photographic processes or one specifically focused on the work of Frederick Hollyer would not rate the match to loc92512466 so highly. Therefore whilst it would be possible to calculate the precision rates of the approach in the case of perfect matches, it would be difficult to account for semi-referent results given their subjective nature.

7.1.1 Data collection

Due to the lack of a suitable gold standard dataset against which the proposed approach could be run, there was no way to produce an absolute measure of

performance[10]. Therefore, it was decided that the best approach was to conduct a comparative analysis. This measured the relative performance of the proposed approach compared to that of manually searching. A primarily quantitative, mixed methods approach was chosen, based around a survey to be completed by the test participants. Quantitative analysis was conducted on closed responses which rated the relative performances of the proposed and manual search approaches. If quantitative analysis proved insufficient then limited qualitative analysis could be conducted on open responses collected during the same survey. This method was selected as there was a concern that there were not enough participants available for testing to allow for a purely quantitative analysis of the approach. Quantitative analysis was the preferred approach as this allowed easy comparison between the two approaches and potentially to co-reference identification systems employed in other domains. However given the limited number of participants and, therefore, responses available, qualitative questions to record the participants attitudes towards the proposed approach and therefore provide a degree of contextualisation for the quantitative data collected from them were included. This is, therefore, a study with quantitative priority and analysis of subjective quantitative responses[167] with additional qualitative analysis conducted to reinforce the quantitative discoveries if necessary. Under the six mixed methods design strategies as described by Creswell, the chosen approach was a concurrent nested strategy. I.e. priority was given to one method while another is contained within. The aim of the nested method being to address different questions to the priority method. In this case aim of the nested, qualitative method is to identify potential differences in search style and attitudes towards the proposed approach between the participants as opposed to the result quality/recall focus of the quantitative analysis[11].

Selection and recruitment of test participants was the first issue which needed to be addressed. As this research has been conducted so as to assist photo-historians and not for general image searching, the decision was made to restrict the testing participants to those individuals with at least some prior knowledge and experience in the area of photo-history. Whilst this significantly reduced the number of potential participants available, and therefore precluded a fully quantitative analysis approach, the increased expertise with regards to searching museum collections portals and knowledge of photographic terminology and processes was considered

[10]I.e. benchmarking.

[11]See section 7.1.2 on page 138

necessary. The search approach used is affected by familiarity with and knowledge of a domain[95, 109, 233]. Therefore, the searches of individuals with expertise in photo-history would be different to those of the general public, and the results they found would likely be different.

Rubin[187] suggests a minimum of four participants for user testing when attempting to identify significant problems. He recommends at least eight participants if a single test is being conducted as opposed to an iterative series. Krug[114], however, recommends three and no more than four participants. Given the mixed methods approach employed, and the desire to employ some quantitative analysis it was decided to have a target of at least eight participants as the minimum sample size.

Overall eight individuals participated in the testing and searched for a total of twenty two different records from the ERPS collection. All possessed a strong interest in and had experience with photo-history collections, records and images. The raw ranking results collected during the testing can be seen in table F.2 on page 216.

Survey data collection was conducted over a period of two months. Test participants were asked to select three records[12] from a list of 795 records selected from the ERPS collection[13]. The participants were asked to try and find matches for the records that they had selected in the six collections[14] included in this research. They were allowed and encouraged to use any search engines, external knowledge and additional resources that they desired but advised that only matches in these six collections were of interest. In addition to the three records that each participant selected, all participants were asked to search for an additional two pre-selected records[15].

[12]For the purposes of testing the proposed approach, the best case scenario would have been for every test participant to examine the results for every test record. In reality this was not possible. Three user selected records were settled up as an acceptable compromise between the desire for as much test data as possible and the willingness of the test participants to assist in this research. Fewer records may not have provided sufficient test data from which to draw conclusions, more would have been an excessive imposition on the time of the test participants. As it was, not every test participant conducted the full five (three participant selected plus two pre-selected) searches that were requested, due to the time involved.

[13]The 795 are a subset of the 1,040 ERPS records with images. The reasons that only 795 of the 1,040 ERPS records with images were available for testing are discussed in section 7.1.3 on page 146.

[14]BkM, DNZ, ERPS, LoC, PEiB and V&A.

[15]erps17093 and erps28409.

Given that searching, even for five records, can talk a considerable amount of time depending on the thoroughness of the search; the test participants searched independently and outside of the lab. Primarily this decision was to make to ensure that the testing process was as convenient as possible and have as many of the contacted individual complete the searches as so become test participants. Another secondary consideration was allowing the participants to search as naturally as possible, this meant wherever and using what browser/operating system combination they wanted. Providing access to this within a lab setting would have been very difficult. From a testing quality perspective, however, a more controlled search environment would have been preferred as this would have presented additional data gathering opportunities (i.e. exactly how the participants searched across the collection) and would have resulted in a greater confidence in certain information (i.e. time spent searching) rather than relying on self reporting. A basic understanding of search behaviours demonstrated by photo-historians has already been acquired through informal discussions as already mentioned in section 1.2 on page 8 but a more in depth understanding and analysis could have proven beneficial. In the end, however, it was decided that the potential benefits of a more controlled search settings would not compensate for the significantly increased difficulty and inconvenience for the test participants and which could have resulted in a significant lowering of the response rate. Given the existing difficulties in recruiting suitable individuals this was unacceptable.

The records which the participants selected themselves allowed the participants to search for records and/or topics which interested them and/or which they had prior experience with. It was hoped that the participants would search more thoroughly for records which they had personally picked and this would, therefore, give manual searching its best chance.

Since the participants were unlikely to select the same records, this also increased the total number examined during the testing. Given the anticipated low occurrence of co-referent matches in the collections, the pre-selected records ensured that the participants would be conducting searches when co-referent matches (of varying degrees) were already known to exist. As the same two records were searched by all participants, this also offered the possibility of calibrating the results of the individual participants relative to one another. Whilst a larger number of pre-selected records would have been preferred, as this would have provided a wider range of records where an in depth analysis of the match scores could have

been conducted, this would have further reduced the breadth of records examined by the testing participants. In order to ensure a reasonable degree of freedom for the participants to search as they saw fit and to allow for a larger proportion of the 795 processed records to be examined, only two records were pre-selected.

The two ERPS records selected for all the participants to search for were erps17093[16] and erps28049[17]. erps28049 was known to have highly referent matches in both the LoC and V&A collections. The best match known for erps17093 was fairly poor in comparison. In this case, the match was located in the ERPS collection and consisted only of a photograph of daffodils[18] exhibited by the same individual a year prior. It should be noted that a simple keyword search using the terms found in either of the *title* fields from either of the two records shown in table 7.2 on the following page would not have found this connection. Therefore, the erps28409 searches were expected to produce higher match ranking values than erps17093.

Having completed their manual searches, each participant asked to complete a questionnaire, either in person or over Skype . During the questionnaire the dendrograms produced by the proposed approach[19] for the ERPS records that the participant had searched for were presented. The participants were asked to compare the results found via the proposed approach to those found by their manual searches. The dendrograms were only made available to the participants after they had completed their manual searching as making them available prior to this could have affected their manual searches. Either by indicating that there were co-referent records available to be found and where, or by leading them to believe that there were no matches to be found when a manual search would have been successful if attempted.

The questionnaire followed a concurrent data collection strategy[20][44, 56, 212] for the collection of the quantitative and qualitative responses to the result comparisons. In this case a comparative quantitative results of the manual and proposed search approaches is collected for one set of search results followed by the potential for qualitative responses for the same search results. The approach

[16]See table 7.1 on page 132.
[17]See table 7.1 on page 132.
[18]The search record was of "pheasant-eye narcissus", a type of daffodil.
[19]See figures 6.13 on page 125, 6.14 on page 126 and 6.15 on page 126 for an example.
[20]I.e. the quantitative and qualitative data is conducted at the same time to validate one another.

Id	erps17093	erps16640
Title	Pheasant-eye Narcissus	Fair Daffodils
Person	H. T. Malby	Henry Thomas Malby
Process	Platinum (Print)	Bromide (Print)
Date	1896	1895
Image		
Found by	N/A	Test approach
Attribution	Copyright ©2008 De Montfort University. Database right De Montfort University (Maker). All rights reserved.	

Table 7.2: Co-reference candidates for erps17093. Although the two records are not of the same photograph, they are clearly related, having been taken by the same photographer and exhibited just one year apart.

then moves on to collect the responses to the next set of search results where the quantitative and qualitative collection repeats. This approach repeats until all sets of search results have been compared. This strategy has the advantage of being intuitive and easy to understand for the participants but preclude follow-up investigations into interesting responses. It is, however, a proven approach[56].

During the search result comparision the participants were asked to ignore the appearance of the dendrogram from a design/user interface standpoint and focus instead on the ordering and relevance of the results. The graphical appearance of the returned results was a usability issue, and was therefore not a focus of the interviews. The participants were allowed to investigate as far down the dendrogram as they wished, follow whatever connections and examine any results

they wanted.

Rankings for result quality were recorded by the survey in two ways. Firstly as values on an eleven point scale of closed responses with a midpoint, ranging from "No relevance" to "Found a perfect match"[21]. Secondly, the participants recorded which set of results they preferred, both at an individual record level and also for all records compared. Preference was recorded on a five point scale of closed responses with a midpoint[22].

7.1.2 Analysis

Given that the performance of a co-reference identification or search system can be analysed as the combination of two separate factors (recall and precision, see section 2.), the testing analysis needs to consider each factor separately before considering what the combined performance is.

7.1.2.1 Recall

As discussed in the previous section, the majority of searches were not expected to find matches of any significance. However this was expected to be the case for any and all search approaches. The point of interest was which approach was most effective at finding co-reference candidates when they do exist. In other words, it needed to be shown that the proposed approach was not missing co-reference candidates that a manual search would otherwise have found.

During testing, matches were found for seven out of the twenty two searched for records. Whilst the strong co-reference candidates for records erps17093 and erps28409 were known in advance of the testing, the potential matches for erps16545, erps16578, erps16939, erps18912 and erps18559 were not. The potential matches in this case were identified as those records which were given a rating of > 5 by the test participants[23].

Figures 7.2 and 7.3[24] show breakdowns of the proportions of matches which were

[21] See section I.2 on page 231.

[22] See section I.2 on page 231.

[23] I.e they were in the top half of the survey scale and therefore considered by the participants to be more matching than not. The scale was deliberately designed to resemble a Likert scale, a proven approach in satisfaction surveys[124].

[24] See table F.1 for raw data.

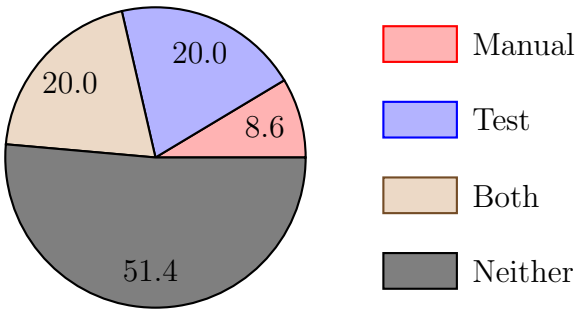


Fig. 7.2: Percentage of tests in which each approach was deemed to have found a co-reference match.

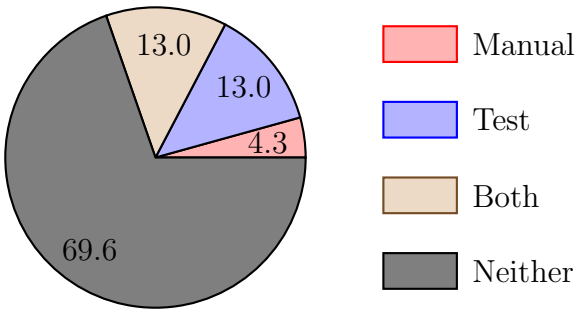


Fig. 7.3: Percentage of distinct records for which each approach was deemed to have found a co-reference match.

identified by manual searching, the proposed approach, by both techniques and by neither technique. The different figures show the proportions on a per record and participant basis and a per unique record basis. To clarify, both erps17093 and erps28409 were tested eight times during the course of the testing. In figure 7.2 each of those tests constitutes a separate result, however in figure 7.3 those sixteen tests only constitute two results as only two distinct records were tested.

As the figures show, the proposed approach found some form of match for 26% of the searched for records while manual searching achieved the same for 17.3% of searches. There was some concern over the inclusion of the results from the preselected records in the recall analysis. As these two records were already known to produce co-reference matches using the proposed approach prior to testing, the concern was that this might constitute an unfair advantage over manual searching. However, if the results of the two preselected records were excluded from the analysis, the recall performance of manual searching relative to the proposed approach was reduced further, achieving matches in only 9.6% of searches compared to 19.1% for the test approach. With the preselected record results included, manual searching achieved a recall rate equivalent to 66.5% that of the proposed

approach, with the results excluded that drops to 50.3%. Therefore, it was clear that the proposed approach outperformed manual searching in terms of result recall.

Also of interest was the number of times that only one of the two approaches found a match. One unlikely but possible outcome would have been for each search approach to find completely distinct and mutually exclusive matches, each approach missing a significant number of matches that the other found. As figure 7.3 on the preceding page shows, this occurred in $\geq 17.3\%$ of cases. Whilst this suggested that there was still room for improvement, overall the proposed approach offers a significant improvement (in terms of recall) over manual searching, missing matches in just 14.1% of cases which produced a match compared to 42.9% for manually searching.

Probably the most interesting searches were those occasions when only one of the tested approaches successfully found a match. These examples provide the clearest examples of the failings for each of the approaches. For example, erps16578 was successfully matched to a record[25] from the BkM by participant 3. However this record was not located by the proposed approach. Subsequent investigation has revealed that although the record in question was available via the website, it did not appear to be available via the BkMs REST API. Following communication with the BkM it was determined that only a portion of the records available via the website are accessible via the REST interface and this records was not one of them.

In another example, when searching for erps16939 using a manual search the testing participant failed to find erps22432[26] despite both records having identical *title*, *person* and *process*[27]. Only the *date* field changed from 1896 to 1903. This demonstrates the issues arising due to the fragmented nature of the search space. The test participant in this case simply did not include the ERPS collection as part of their manual searching. It should be noted that all participants were been supplied with a list containing all six collections of interest, including ERPS.

7.1.2.2 Quality

The match quality was a separate issue to that of the match rate. In the case of erps28409, for example, different participants found different potential co-reference to matches for it. Although a large number of the testing participants found a co-

[25] See table G.2 on page 218.
[26] See table G.3.
[27] Although the *process* fields just contain "[Not Listed]".

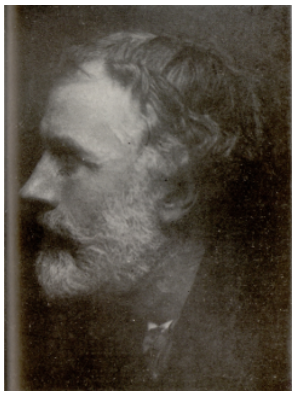
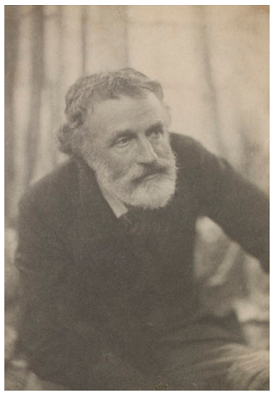
Id	erps28409	vaO101352
Title	George Meredith	George Meredith; Portraits of many persons of note photographed by Frederick Hollyer
Person	Fred. Hollyer	Hollyer, Frederick
Process	[Not Listed]	
Date	1909	1886
Image		
Attribution	Copyright ©2008 De Montfort University. Database right De Montfort University (Maker). All rights reserved.	©Victoria and Albert Museum, London.

Table 7.3: Example of a promising but non-matching co-reference candidate for erps28409.

reference candidate for erps28409 and the recall rate for erps28409 was, therefore, quite high[28], those co-reference candidates could have been significantly better (or worse) candidates than the records found by the proposed approach. Therefore, the quality of the co-reference candidates found by the two approaches had to be compared in order to produce a fair assessment of the relative performances of the two techniques.

This was done in two ways. Firstly the results of the two preselected records were analysed to determine if the distribution of ranking values for the two approaches differenced significantly. A statistically significant difference would suggest that one of the two approaches was producing better results than the other, whilst a lack of significant differences would suggest that when both approaches find matches they are of equivalent quality. This analysis could only be conducted on the two preselected records as none of the other records were examined by enough participants for a distribution of rankings to be available.

Secondly, the participants' preference responses were examined in order to

[28]87.5% of participants found a co-reference candidate.

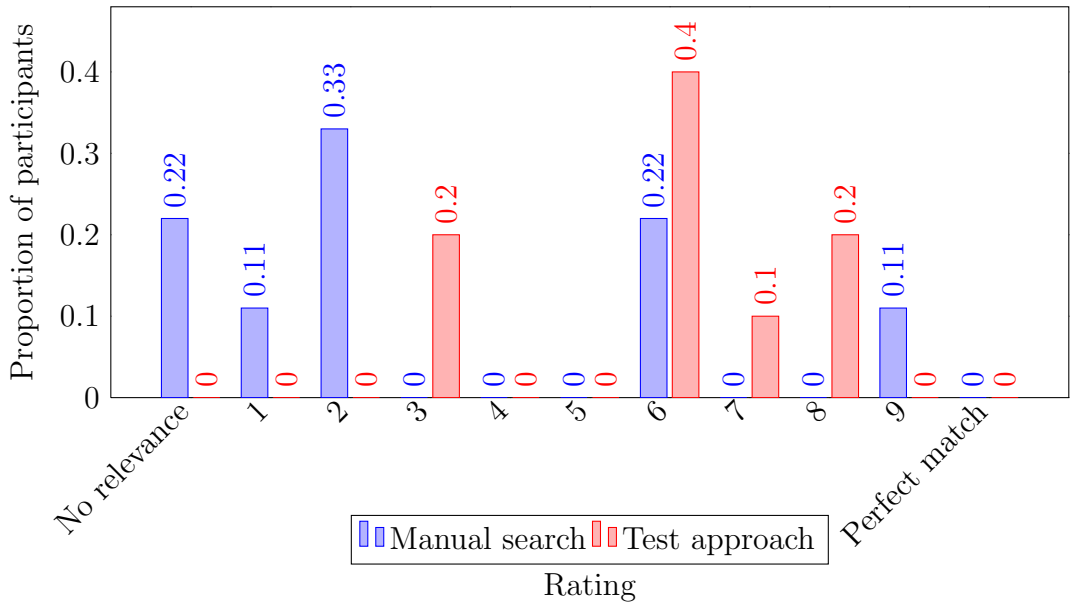


Fig. 7.4: Match quality ratings for erps17093 given by the testing participants. Note that although the similarity values are spread across the range, the majority of results for the test approach are grouped towards the higher end of the scale.

produce a more qualitative measure for how well the participants felt that the search approaches were performing relative to each other.

The participants' ratings of the matches produced by the proposed approach for erps17093 ad erps28409 are shown in figures 7.4 and 7.5 respectively.

The small number of test subjects available meant that many statistical analysis techniques were unsuitable[29] Therefore, in order to test the significance of the apparent difference between the results the Mann-Whitney U (MWU) test was used. The MWU test is a non-parametric approach suitable for testing two independent data sets[135]. Valid use of MWU rests on the following assumptions[201]:

1. Dependent variable - Must consist of ordinal or interval values. For this test, the dependent variable are the ordinal co-reference ranking values.

2. Independent variable - Must consist of two independent groups. For this test, the two groups are the search approach used.

3. Independent observations - The ranking values for manually searching can not be affected by the ranking values for the new search approach.

[29] I.e. *t*-test.

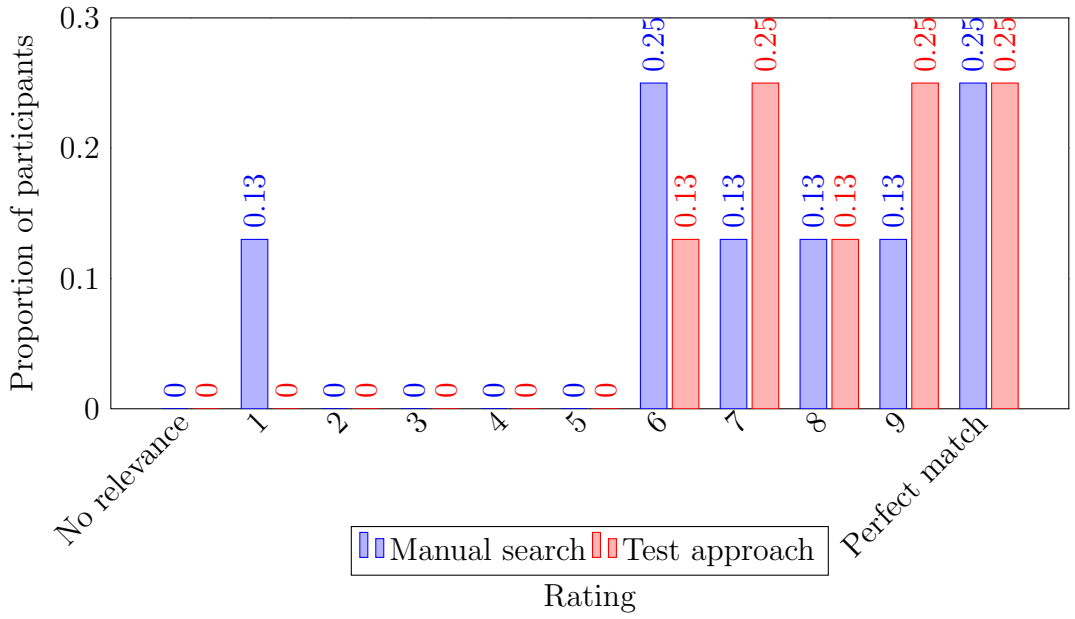


Fig. 7.5: Match quality ratings for erps28409 given by the testing participants. Note that the results are not as distributed as those seen in figure 7.4 on the preceding page and are more strongly grouped towards the higher end of the scale. This is in line with the prediction made in section 7.1.1 on page 132 that matches for erps28409 would be easier to locate than those for erps17093.

Importantly MWU has dramatically lower sample size requirements when compared to other approaches[30]. In addition to the low sample size requirement for MWU, it can also be used when the values being compared are not normally distributed. Whilst the participants' responses are likely to follow a normal distribution, this was not tested for given the capabilities of the MWU test.

Using MWU, the hypotheses were as follows, H_0 = there was no difference between the two sets of results and, therefore, the results of manually searching and the proposed approach are equivalent. H_1 = there was a difference between the two sets of results and, therefore, one approach performs better than the other. If the null hypothesis was disproved[31], then further analysis would be required to demonstrate which approach was better performing. However, based on the mean average rankings for the records, the hypothesis would be that the proposed approach was the better performing.

IBM Statistical Product and Service Solutions (SPSS) was used to perform the MWU tests and the results can be seen in table 7.5. As the results show, the null

[30] I.e. *t*-test.
[31] H_1 was true.

Record	Approach	N	Mean Rank	Sum of Ranks
erps17093	Manual	8	6.38	51
	New	8	10.63	85
erps28409	Manual	8	7.69	61.50
	New	8	9.31	74.50

Table 7.4: Average rankings for the pre-selected test records.

	erps17093	erps28409
Mann-Whitney U	15.000	25.500
Wilcoxon W	51.000	61.500
Z	-1.818	-0.694
Asymp. Sig. (2-tailed)	0.069	0.487
Exact Sig. [2*(1-tailed Sig.)]	0.083	0.505

Table 7.5: MWU test results.

hypothesis (H_0) was not disproved for the results of either erps17093 or erps28409 (p values of 0.069 and 0.487 respectively[32]). There was, therefore, no significant difference in the distribution of values for either the manual or proposed approach.

As predicted[33], the erps28409 results produced higher average rankings than those of erps17093 and in both cases the mean and sum of ranks values for the new approach results were higher than those of manually searching (see table 7.4). However the results of the MWU tests do not show statistically significant differences in the ranking values produced by the two approaches. Therefore, it seems that the results of manual searching and the proposed approach are equivalent in terms of the quality of co-reference candidates that they find, for the two pre-selected records at least.

Whilst the actual performances of any search system is important, if the approach is perceived to produce lower quality results then users will simply not use it regardless of any objective success rate. Therefore, in addition to the statistical analysis of the preselected records, the test participants were also asked which approach they thought produced the better results as a qualitative measure.

Figure 7.6 shows the participants responses when asked which set of results they preferred on a per record basis. The responses including and excluding those for the pre-selected records are shown. The responses show a slight preference for the

[32]Statistical significance requires at least ≤ 0.05.
[33]See section 7.1.1.

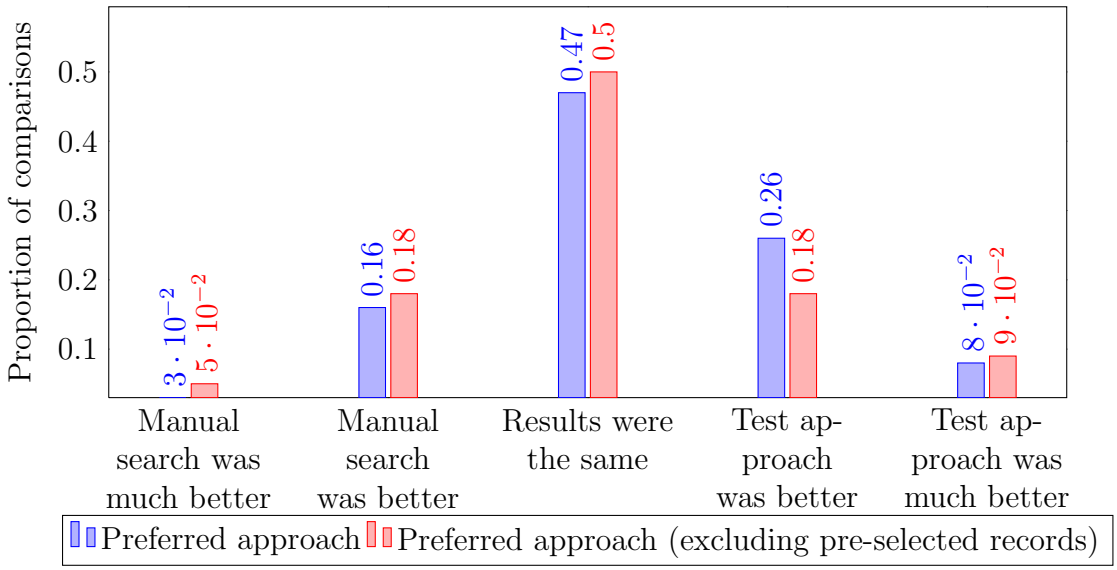


Fig. 7.6: Approach preferred by test participants on a per search basis.

results of the proposed approach in both cases; however, the difference was too small to be considered proof of preference. Therefore, it again appears that the results of manual searching and the proposed approach produce co-reference candidates of equivalent quality.

Further evidence for the effectiveness of the proposed approach comes from the final survey question. In which the participants were asked if they would consider using the proposed approach if it was made available to them. As the results in figure 7.7 show, the majority of participants stated that they would use the new approach. Half of all participants said that they would definitely use it, and no participant said that they would not. The 'maybe' response was given by participant 7. As the participant with the greatest level of experience in photo-history, 7's search style had certain differences to those of the others. Specifically, due to their experience in the field, 7 already knows the most promising location/collection for many photographs based solely on photographer etc. Therefore, they were able to go straight to the relevant collection instead of conducting a more exploratory set of searches.

That the search style used would be affected by prior domain experience and knowledge was expected, although the precise effect was not known. Although studies have been conducted to investigate the effect previously, the results and conclusions of these studies have been inconsistent and, arguably, incompatible with each other[200]. The difficulty in determining the effect of prior knowledge may lie, in part, in the difference between having prior knowledge of a specific

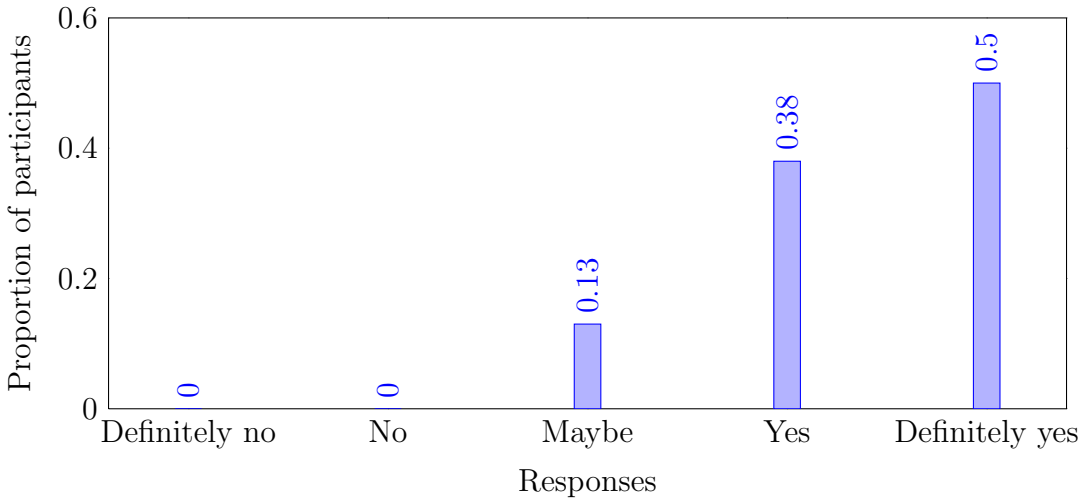


Fig. 7.7: Responses to the question "If the test approach was made available to you, would you use it for searching in the future?" as percentages.

domain and prior knowledge of the, or similar, search tools. While older individuals can have greater knowledge in certain domains, younger individuals have greater familiarity/knowledge with online search tools[36]. That prior knowledge has an effect is clear, exactly what that effect is appears to be dependant on multiple factors, including interface and subject domain. The results of this testing and other investigations[36] suggest that, for photo-history at least, the effect is to produce a more directed search with less exploratory queries. Further research would, however, need to be conducted to confirm this hypothesis.

The conclusion that can be drawn from these results is, therefore, that the proposed approach produces results of equivalent quality to those of manual searching. However, there was a suggestion that, as photo-historians gain experience with GLAM collections, their need for a search system of any kind may be reduced due to their domain experience.

7.1.3 Testing problems

When processing the ERPS records, practical difficulties arose which required manual intervention in order to produce the final result dendrograms. These difficulties arose because the co-reference identification approach described in this thesis was very memory intensive. In order to produce the final dendrograms, the computing hardware used needs to be able to calculate and store the similarity matrices for

title	*person*	*process*	*date*
0.831	0.080	0.022	0.216

Table 7.6: Average number of unique values per record.

the individual fields, as well as the resulting overall similarity matrix. However, the memory required to store these matrices could exceed the memory available for this research project under certain circumstances. Unfortunately, these circumstances were met for a large proportion of the ERPS records[34].

Figure 7.8 shows the predicted size of the combined individual field similarity metrics compared to the number of records being compared. The size of the field similarities matrices could be calculated perfectly if the exact number of unique field values were known using equation 7.1. n represents the number of unique values and b the amount of memory required to represent the similarity value[35].

$$\text{mreq}(n) = b \cdot \frac{n^2 + n}{2} \tag{7.1}$$

The number of unique a values in each field can only be found by actually counting them, however the value can be predicted with a high degree of accuracy. Table 7.6 shows the average[36] number of unique values found for each field per record being compared[37]. The total size of the individual field similarity matrices can, therefore, be estimated as shown in equation 7.2. r represents the number of records being compared. A comparison between the predicted and actual memory requirements for various ERPS records can be seen in figure 7.8.

$$\text{mreq}(0.831r) + \text{mreq}(0.080r) + \text{mreq}(0.022r) + \text{mreq}(0.216r) \tag{7.2}$$

As expected, the size of the combined field similarity matrices increased as more and more records were compared. Therefore, the amount of space needed to store the overall record similarity matrix also increased[38]. As shown by figure 7.9, there comes a point where the amount of memory required exceeds the memory available, assuming 7GB as the total memory availability[39], this point was reached

[34]See figures 7.8 on the following page and 7.9 on page 149.

[35]For this research C++ floats were used to store the similarity values, therefore $b = 4$ bytes.

[36]Mean.

[37]Based on an analysis of the actual number of unique values found across 795 different searches.

[38]The overall record matrix size follows a similar progression to that of the individual fields, however since records do not need to be compared to themselves the equation was $b \cdot \frac{n^2-n}{2}$.

[39]The hardware used for this research had a total of 8GB but a limit of 7GB used.

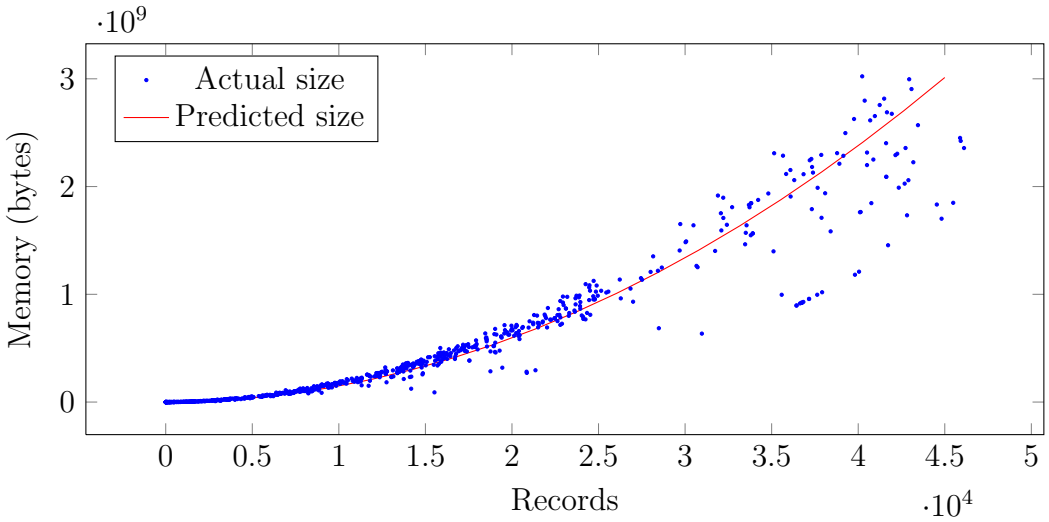


Fig. 7.8: Combined field similarity matrix sizes verses predicted size.

at ≈ 16,500 records. Real world deployment of this approach would therefore require, either a machine with better hardware specifications or changes in the process so has to lower the memory requirements. The 7GB limit is, therefore, not an absolute limit which would prevent real world deployment, it only restricts the records which could be used for testing. As this project was intended to determine if matching photographic museum collections could be achieved at all and was not intended to be a real world test of software, the subset of the 1,040 that the 7GB limit left available[40] still allowed the test participants to select from a wide range of records covering many different topics.

Fortunately it was not necessary to store the full record similarity matrix in order to generate the similarity dendrogram. A large number of the similarity values in the overall record matrix will be at or near zero. These links were of very little value since they described extremely tenuous connections between records. By sorting the overall record matrix by similarity values and taking only the top t percent of the ordered matrix, the near zero values are discarded which massively reduces the space required to store the matrix. Discarding values in this way can and does mean that some records will not be successfully linked to the similarity dendrogram, but this only applies to the records with the lowest similarities compared to all other records and which are, therefore, of little or no interest. The value of t was determined by the combined size of the individual similarity matrices and the number of records being

[40] 795

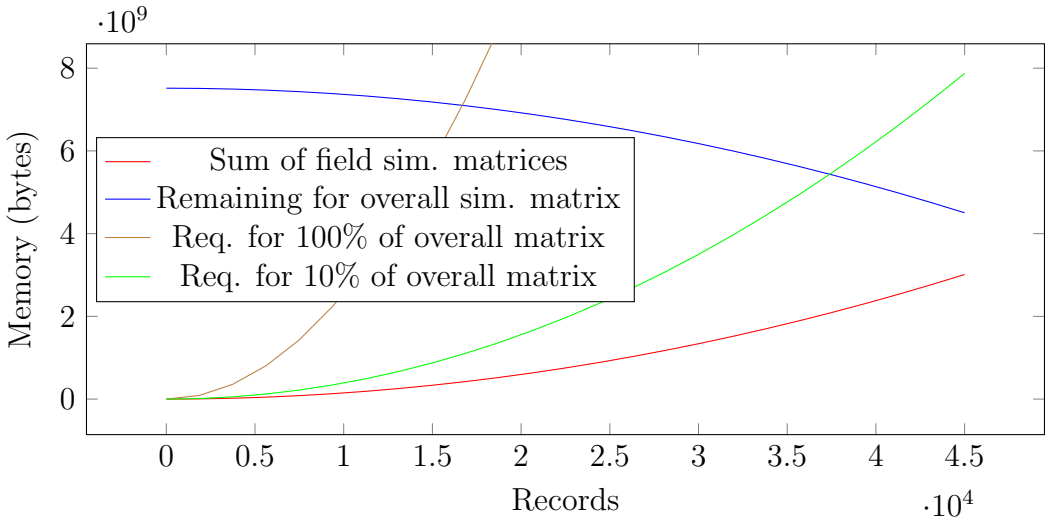


Fig. 7.9: Memory requirements for proposed approach.

compared. For the searches a limit of $t \geq 0.1$ was set, if t fell below this value then the dendrograms produced would be overly affected. Figure 7.9 demonstrates that storing only 10% of the overall similarity matrix allowed significantly more records to be processed before the memory requirements exceeded the available memory.

As figure 7.9 shows, it was possible to store the full overall record similarity matrix when comparing fewer than ≈17,100 records[41]. The maximum number of records that it was possible to compare with the 7GB limit in place was 33,098 where it was only possible to store 10.109% of the overall matrix. The number of records that could be compared whilst respecting the t limit was somewhat flexible. However, as the size of the overall similarity matrix which can be stored was dependent on the space left over from storing the individual similarity matrices. If there were a large number of duplicate field values in the records being compared then the individual field similarity matrices were small, leaving more space available for the overall similarity matrix.

7.1.3.1 Keyword filtering

Unfortunately, even when storing only the top 10% of the overall record similarity matrix there were still a large number of ERPS records which could not be processed given the available resources. With the limit in place, only 49.1% (511) of the 1,040 records could be processed. The problem was that too many records were being returned by the initial query expansion phase of the proposed approach. In

[41] 17,101 records = 100.058%, 17,116 = 99.7952%

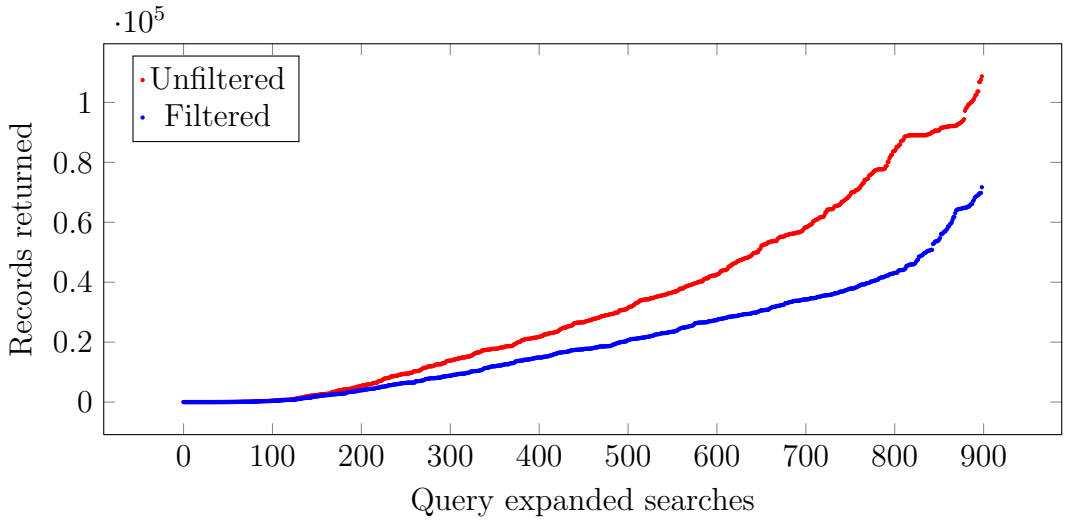


Fig. 7.10: Distributions for the number of records returned with and without keyword filtering.

order to increase the number of records available for testing, the lists of query expanded keywords for the records with t values of < 0.1 were manually pruned and the searches conducted again. Pruning the keywords reduces the number of results returned by the keyword searching of the various external GLAM collections and so reduces the size of the similarity matrices. However, care had to be taken when selecting keywords for removal. Only those keywords included due to topic drift by the query expansion processing were removed. When it was not possible achieve a t value of ≥ 0.1 without removing valid expanded keywords or keywords which appeared in the *title* or *description* fields directly, those records were excluded from the testing. Processing of those records would require either increasing the available memory (which would have been a temporary fix and would only delay the appearance of the issue) or lowering the minimum accepted t value (which would eventually produce significant reductions in the quality of the record ordering in the dendrograms).

The results of filtering the query expanded keyword limits on the number of records returned compared to unfiltered searching are shown in figure 7.10. Following the pruning, 76.5%[42] of the 1,040 attempted records were successfully processed.

[42] 796

7.1.3.2 Discussion

Manual trimming of the query expanded terms was an unfortunate but necessary step. Trimming the keywords in this way would not be acceptable in a real world version of the proposed approach. It would, therefore, need to be revisited as part of future research. However for testing purposes it was felt that the additional records which trimming the search terms made available for testing, justified the manual assistance which was required in order to process them.

The issues regarding the size of the similarity matrices are a significant barrier to further expansion of the approach. As the number of the collections increases, the number of records returned during the initial query expanded keyword searches will only increase. Already the number of records returned from just six collections was shown to cause problems. Whilst the issue can be mitigated/delayed by using more powerful computers, as matrix size increases quadratically, the amount of storage space required will increase very rapidly. If this research were taken further, then methods of reducing the number of records which need to be compared or of reducing the size of the subsequent similarity matrices need to be investigated.

7.1.4 Conclusions

Overall the results demonstrate that the results found by the proposed approach are at least equivalent to those produced by manual searching. The recall of the proposed approach is better than that of manual searching, and no statistically significant difference was seen in the quality of the matches found by manual searching or the proposed approach. However the problems encountered in processing the records used during the testing demonstrates that further refinement of the proposed approach is required because it can realistically be used.

7.2 *Title* metric testing

In addition to the result recall rates and result quality, the *title* metric was judged to require specialised testing. The *title* field could have been addressed using existing techniques[43][52, 123]. These approaches were not used since they were viewed as being too computationally expensive and instead a custom and computationally

[43]I.e. LSA or STASIS.

lightweight approach was created. The quality of the results produced by the new *title* metric was predicted to be lower than the results produced by more established techniques since the new approach was computationally simpler; however, the decrease was expected to be acceptable given the significant gains computationally. The suitability of the novel *title* metric used in this research was, therefore, dependent on the new approach being measurably faster.

The *title* metric was tested in order to determine the quality of the textual similarity values it produced and the time it requires to produce its results. The aim was to determine firstly, if the quality of the similarity results produced differed between the *title* approach and the established techniques. Secondly, if the time needed to calculate their results differed significantly between the *title* metric and the established approaches and if the time difference was sufficiently significant to mitigate the anticipated drop in result quality.

7.2.1 Data collection

In order to measure the quality of the textual similarity values produced, the results of the *title* metric needed to be compared to those of established approaches. The performance of LSA and STASIS have already been compared by O'Shea et al.[164]. In their paper, the results of the two approaches are compared to the averaged similarity scores from human testers[44] using a subset of the STSS-65[123] dataset. Given the availability of an existing set of results with existing gold standard results to compare against, the simplest way to measure the quality of the results from the *title* metric was run it using the same testing data as used in the work by O'Shea et al. and directly compare the values produced against those reported for LSA and STASIS. Therefore, the *title* metric was run against the same STSS-65 subset used by O'Shea et al.[45].

The testing data (STSS-65) consists of word pairs[46]. Whilst it more closely resembles the contents of the *title* fields than many other data sets[47], it was not a perfect emulation. As such, the results produced are only approximate representations of the relative performances of the tested techniques if they were run on the contents of *title* fields. Figure 7.11 shows the LSA, STASIS and human produced

[44]Human testers are considered to produce the best results when measuring textual similarity.
[45]See table B.1 on page 207.
[46]See table B.1 on page 207.
[47]I.e. Microsoft Paraphrase[181].

similarity values plotted with the results from the *title* metric.

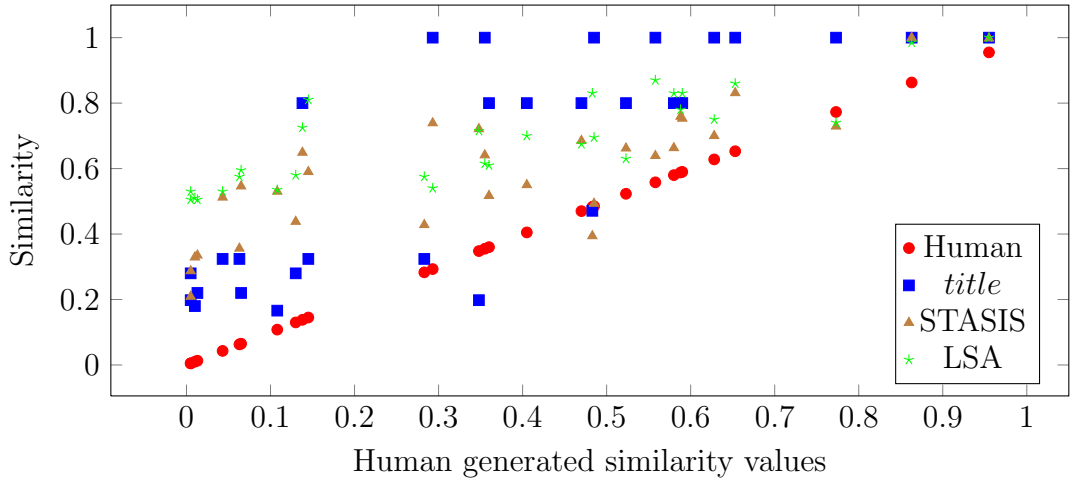


Fig. 7.11: Human, LSA, STASIS and *title* metric generated similarity values for STSS-65 subset.

The computational requirements testing was conducted using Python implementations of all three approaches running on an Intel Core2 Duo T5500 (1.66GHz) machine. Alternative programming languages and/or hardware were known to result in faster implementations, but as testing was intended to demonstrate the relative performances, the absolute performances were unimportant.

Five sets of results were produced, in each one the time recorded was the time taken to produce a non-directional pairwise similarity matrix for the *titles* being compared. The *title* fields used for the testing were a random selection from the 1.7 million records collected as part of this research[48]; however, the same random selection was used for all the tests. The first shows the time taken by LSA. The second shows the time taken for the *title* metric to do the same using pre-calculated word similarity values. The third shows the time taken for the *title* metric if the word similarity values are not pre-cached. Since each word pair needs only be compared once and can then be stored in perpetuity, starting with no pre-cached word similarity values would be unlikely, these results are, therefore, included only for completeness. The fourth shows the time taken by STASIS using pre-calculated word similarity values. The fifth shows the STASIS time without pre-cached values. It should be noted that the *title* metric and STASIS have different methods for calculating word similarity values and the appropriate approach was used in each case.

[48]See section 6.2.1.

7.2.2 Analysis

Following on from the approach used in the paper by O'Shea et al.[164], the *title* metric values were compared to those of the human responses using Pearson's correlation coefficient. Pearson's was used as it is the standard statistical technique for measuring the linear relationship between two variables. In this case, the first variable was the human response and the second was the response from the similarity metric being examined. The results of the *title* metric produced a correlation value of 0.807 compared to 0.838 for LSA and 0.816 for STASIS. This means that the *title* metric represents a performance decrease of just 3.699% compared to LSA.

A standard Z-test can be used to compare the correlation coefficients for any two of the approaches. A Z-test can determine if there is a statistically significant difference between them given the sample size used to produce the correlation values. However, a Z-test assumes that the values being compared are normally distributed, which is not he case here. Fortunately this can be resolved using a Fisher Z-transform to transform the correlation values before comparison[68, 176]. Comparing the correlation values for LSA and the *title* metric in this way produced a Z value of just 0.35 where the Z critical values are 1.96 for $p < 0.05$ and 2.58 for $p < 0.01$. Therefore, there is no statistically significant difference between the coefficients of the two approaches and consequently, no significant decrease in result quality of the *title* metric compared with LSA and STASIS.

Despite the comparable performance of the *title* metric from a quality standpoint, the main justification for the use of this new approach compared to an existing technique was the time taken for processing.

Figure 7.12 shows the time taken for the three approaches. As can be clearly seen, the *title* metric was significantly faster than LSA when using pre-cached results. The performance without pre-cached values was initially worse than that of LSA but quickly improves as the number of records being compared increases. This was because the number of word similarity values which need to be calculated was directly related to the number of unique words in the records being compared. However the number of unique words per record was inversely proportional to the number of records being compared. Therefore as the number of records to compare increases, the proportion of time spent generating word similarity values decreases.

The proposed approach also requires significantly less time than STASIS in both the pre-cached and un-cached tests. Whilst the difference between the *title* metric

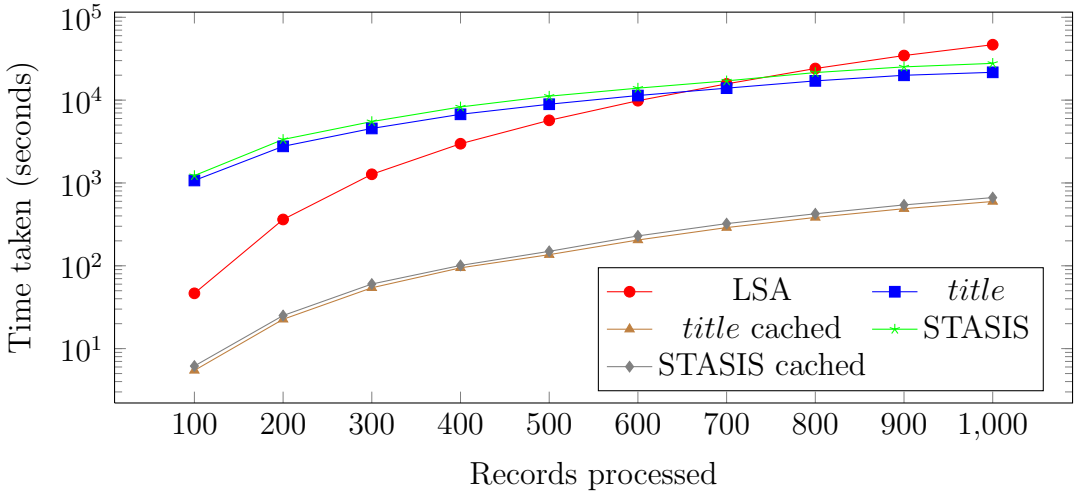


Fig. 7.12: Comparison of processing time requirements for LSA vs. *title* metric vs. STASIS.

and STASIS was less dramatic than between the *title* metric and LSA, the difference was still statistically significant. At 1,000 *title* fields to compare, the *title* metric needs just 79% or 90% of the time required for STASIS for pre and un-cached similarity values respectively. This means that the *title* metric runs noticeably faster than the existing techniques with only a minor drop in similarity value accuracy. The *title* metric is, therefore, an appropriate replacement for more established semantic similarity approaches such as STASIS and LSA in this research project.

7.3 Collections searched

Part of the difficulty in locating the missing ERPS records was due to the widely distributed nature of the search space. The exploratory questionnaire which was conducted[49], suggests that most users will consider between three and five collections when searching. The real searches that the test participants conducted offered a valuable opportunity to confirm the responses to the exploratory questionnaire.

7.3.1 Data collection

The initial questions of the interview focused on the collections that the participants used, both previously and for this test specifically. Whilst the use of the collections included in this test by the participants was almost equal, previous experience with

[49]See section 1.2 on page 8.

the collections was mainly limited to the V&A with a few participants having also previously explored the collections of the LoC, PEiB and ERPS.

7.3.2 Analysis/conclusions

Fully half of the test participants used all six test collections when searching, this was slightly above the number of collections that would be expected to be used given the responses from the search style questionnaire (see fig. 1.8). However, the testing only included six collections in total. As the participants were supplied with a list of the six collections of interest before commencing their searches, it is possible that the participants were simply trying out all of the collections that were available to them and that if searching under normal circumstances, they would use fewer collections. Conversely it is also possible that the participants would have used more collections were they allowed. However, this seems unlikely as during the earlier survey discussed in section 1.2.4 on page 12, the majority of participants reported that they typically search across five collections or fewer. Based on the average number of collections used per participant[50] and the initial questionnaire responses, it seems likely that number of collections examined for this test represents a slight increase on the number of collections normally searched.

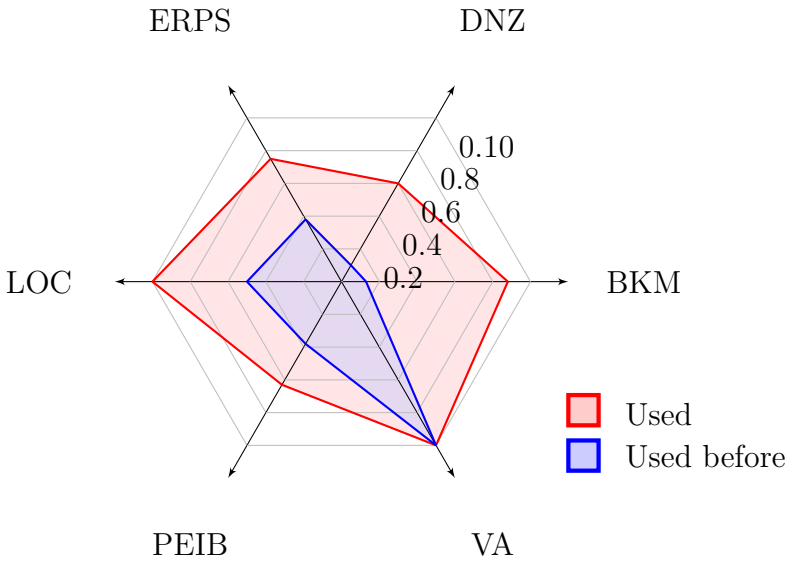


Fig. 7.13: Collections used.

[50]Mean 4.74.

7.4 Time taken

A primary aim of this research was to be able to increase the number of searches that a photo-historian can conduct. This means reducing the amount of time spent searching. In order to know if the new search approach was able to help speed up searches, the current speed of searching needed to be identified. It should be noted, however, that information seeking behaviour in GLAM researchers does have some slight differences to the behaviour of non domain specialists, i.e. the general public. Specifically researchers tend to be more persistent and to search for longer[231].

The test participants were asked how long they spent searching for the five ERPS records overall. The aim was to identify the average time spent searching and so the specific time spent looking for each individual record was unimportant.

On average[51] the participants spent just over ten minutes per search; however, there was a very large degree of variation between participants. Participant 2 spent on average 16 minutes per search whilst participant 3 spent less than $\frac{1}{5}^{th}$ of that time (3 minutes). If the number of collections to be examined increases, the amount of time needed to search through them would also rise. However, as the results of the search style questionnaire and this user testing show, users only search a small set of collections. This suggests that even if more collections were available, they would not be used. Potentially because of the amount of time it would take. In this case, the time that the participants spent searching was not indicative of the time that they needed to spend searching through the collections, but the maximum amount of time that they were willing to spend searching regardless of the number of collections that can be examined in that time.

The time responses also supported the idea that search success is partially dependant on search time. The top 50% of test participants[52] spent an average of 65 minutes searching for an average success rate[53] of 40%. The bottom 50% spent and average of 35 minutes searching for a success rate of 15%. Time spent searching is not the only factor in success. Participants 5 and 8 spent 15 and 35 minutes respectively, however participant 5 had a better success rate at 20% versus 0%.

[51] Mean.

[52] With regards to time spent searching.

[53] Where a match is considered successful if the participant rated it ¿5 on the scale, see section 7.1.2.1 on page 138.

Participant	Time (mins)	Searches	Average
1	60	5	12
2	80	5	16
3	60	5	12
4	60	5	12
5	15	5	3
6	30	5	6
7	40	5	8
8	35	3	11.67
	Average		10.07

Table 7.7: Time spent searching by test participants.

7.5 Conclusions

This chapter described the testing techniques used, results collected and analysis of the new search approach that this research proposed.

When investigating how the recall of the new search approach compares to that of manually searching, the new approach was able to find more potential co-reference matches than manually searching. Out of 23 distinct ERPS records searched for, manually searching found potential co-reference matches for 17.3% while the new approach was successful with 26%.

When investigating how the quality of the co-reference matches found by the new search approach compares to those found by manually searching, the new approach found potential co-reference matches that are equivalent to those found by manually searching. Equivalence between the approaches appeared not just in the absolute performances of the approaches, but also in the test participants' opinions of the approaches.

For the precision and recall analysis, a larger sample size would have been preferred; however, it was felt that having experienced participants (having experience in photo-history) was more important. For an initial evaluation of the suitability of the new approach to the problem, the small sample used was sufficient[63, 187].

When testing the performance of the *title* metric, the results produced by the *title* metric were less accurate than those of either LSA or STASIS when measured against human generated similarity values. However this drop in quality was small.

When investigating how the throughput of the *title* metric compares to that

of established STSS metrics, the throughput of the *title* metric was, as predicted, significantly higher than that of LSA. The benefits when compared to STASIS are less dramatic, but still notable. The *title* metric took ≈88% of the time required by STASIS when using pre-calculated word similarity values and ≈78% when calculating term similarity values on the fly.

Overall the testing demonstrates that the new approach is noticeably more capable than manually searching for finding potentially co-referent records. In the specific cases where co-referent matches were known to exist in advance (erps17093 and erps28049), the new approaches' performance was at least equivalent to that of manually searching, with suggestions that it performed better. Even assuming that the quality of the results was only equal to that of manually searching, the benefits in terms of improved recall would make the new approach a valuable tool. Although the new approach can be slower than manually searching in terms of the total time taken, the amount of time requiring manual interaction would be less since most of the search actions have been successfully automated. Therefore, a user of the new approach can examine more records and search against a wider range of collections than could be achieved in the same amount of time by manually searching.

The potential co-reference examples examined during this chapter were all taken from the 1,040 ERPS records with image data. The reasons for this were that during this testing stage is was necessary to be confident that matches had been located between the searched for and ultimately found records. It is clearly much easier to determine if two records are referring to the same photograph if the photographs are present. The ultimate aim of this research is, however, to locate matches for the ERPS without image data. In acknowledgement of this, section H on page 221 contains potential co-reference matches found for ERPS records without image data. Further investigations into the province of the potential matches would be needed before they are confirmed as copies or originals of the 'missing' photographs, however based on the metadata similarities they appear promising.

One significant problem does need to be addressed, namely the memory requirements. The size of the matrices was expected to be problematic. Certainly for a real world deployment[54] the approach described in the methodology chapter[55] is un-

[54]I.e. Searching for all of the missing photographs using the existing collection and expanded to include as many collections as possible.

[55]See section 6.

realistic without significant upgrades in hardware over that used during the testing. However, the focus of this research was to investigate if it is possible to identify co-referent GLAM photography records at all, not to produce a ready for use system. A variety of methods exist for reducing the sizes of similarity matrices as this is a commonly experienced problem in the area of clustering[213, 224].

8

Conclusions

This research project's specific aim was to discover if it was possible to find the photographs missing from the ERPS exhibition catalogue records[1]. It was hoped that this could be achieved by searching through the contents of photographic collections held by other institutions which had been made accessible over the internet. Whilst this goal could have been achieved by a manual search for the 'missing' photographs, that approach would have been time consuming given the number of records in the ERPS collection (34,197)[2] and the number of disparate GLAM collections which are available to be searched[3]. For this reason, it was hoped that using CI techniques, the 'missing' photographs could be located in either a fully automatic or in a semi-automatic manner. In either case, the person power required to conduct the search would be significantly reduced.

The ERPS exhibition catalogues represent a specific test case in order to demonstrate the issues present in the broader GLAM community and collections as a whole. Namely the imprecise and uncertain nature of GLAM record information and the number of different locations and formats[4] that these records are stored in. The task of searching these collections is complicated still further by the sheer number of GLAM records and consequently the size of the search space.

The research conducted and described in this thesis demonstrates that a semi-automatic approach is possible and as shown by the testing detailed in section 7, is at least equivalent to manual searching in all analysed respects. Whilst the novel search approach described in this thesis is effective at locating highly referent

[1]The important factor is to locate a visual representation of the exhibited images, be that the exact photograph or a copy of it.

[2]See section 3.5 on page 57.

[3]During this research at least 61 institutions were known to have collection APIs or downloadable collections[151], the total number of collections available online in any form is much higher[8].

[4]The CHIN has identified at least 34 separate metadata standards[9].

records, it is not able to make the final determination as to whether or not the results it returns are, in fact, a match for the photographs being searched for. This final stage of confirming the provenance of the photographs described in the records and, therefore, proving the co-reference of the records appears to be an AI complete problem[5] and is, therefore, far beyond the focus of this thesis. A fully automatic approach is, therefore, not currently possible.

Although locating matches for the missing ERPS photographs is a co-reference identification problem, the approach presented in this thesis has broader applications and could easily be modified for use as a generalised information retrieval system[6] for GLAM records. This can easily be achieved by allowing artificial records[7], containing manually entered field values, to be used as the starting[8] record in the proposed approach. Therefore, although the particular focus of this thesis has been locating co-reference matches for photo-history records and the proposed approach address that problem, it could also be used for informational retrieval applications. Furthermore, it seems likely that by changing or substituting the record fields examined and field similarity metrics used; the subject domains and types of records that can be searched by the proposed approach can be opened up beyond photo-history. The most obvious application areas being other sections of museum collection (e.g. paintings) but also domains such as genealogical research where the names of individuals play a major role.

The search approach presented in this thesis makes use of features and techniques from a broad range of areas including query expansion, data cleaning, text similarity algorithms, STSS[9] algorithms, fuzzy logic, path finding and dendrogram generation. The way they have been combined here constitutes an original co-reference identification approach, specifically tuned to the challenges and unique issues of searching for photographic records in GLAM collections. Whilst the approach as described is best suited for searching for records of a photographic nature (the *process* metric in particular), large portions of the search approach

[5]I.e. that solving it would require solving the central AI problem, how to make a computer which is as intelligent as a person?[193].

[6]General search system.

[7]I.e. records were not collected from GLAM collections but which were created based on the search criteria of the searcher. Such records would allow the proposed approach to be used to search using any and all search terms.

[8]Seed.

[9]Short Text Semantic Similarity (STSS), comparing brief pieces of text in order to identify similarities in their meaning.

could be reused if searching for other object types[10] or for generalised searching within GLAM collections or other domains where records are uncertain, incomplete and/or imprecise.

Unfortunately the approach presented in this thesis requires considerable computing resources. This high computational cost comes from calculating the full similarity matrices, both for the individual fields and the overall records. Whilst the time required to calculate these matrices can be addressed by changing the field and overall record similarity metrics, the memory required to store the matrices, which is a result of the number of records being compared, can not. It is the memory requirement of the approach that poses the main problem as it has already caused problems during the testing stage of this research. If the approach was expanded to include more than just the six collections used in testing, then these problems would only be magnified.

Due to the high computational cost of the search approach, the total time required to search GLAM collections using this approach can be greater than the time required for a manual search. However, for the majority of that time, no human involvement is required. Therefore the amount of time the person searching personally has to spend searching is considerably reduced. This means that this new approach meets the initial aims of this investigation in that it can help researchers to locate co-referent images without falling foul of the same prohibitive time requirements of a manual search.

The computational time of the proposed approach could be reduced using more powerful hardware, in particular much of the approach can be run in parallel and so would benefit from multi-core systems. The most significant savings are, however, expected to be achieved from a re-working/writing of the software written during this research to be more efficient[11] and it makes sense for this to be carried out before investing in expensive hardware.

As part of the novel search approach, several effective field comparison metrics were produced, two of which contain significant improvements over existing techniques.

The first of these is the *process* metric. At the present time, searching based on photographic process is limited to keyword searching or selecting from drop-down

[10] I.e. sculptures, painting.

[11] As discussed in section 6 on page 87.

menus in the rare collections where the record's processes have been manually examined and sorted. The combination of graph traversal and approximate string comparison used in this thesis assists in linking photographs even in cases of typographical errors and photographic misidentification. This is not possible using existing approaches. Therefore, this original approach is significantly more capable than those currently employed.

Secondly, the *title* metric represents a significantly faster approach to measuring STSS when compared to established approaches such as STASIS or LSA. Whilst it is less effective at modelling semantic similarity than other approaches when compared to human levels of performance, no statistically significant difference was found during testing. In situations when computing power or time is limited, the *title* metric is an efficient new approach which offers favourable trade-offs compared to existing algorithms.

In addition to demonstrating an effective approach to the central research problem, a logical and initially promising approach was ruled out during this investigation, specifically the use of clustering. Whilst clustering was able to identify valid groupings within the combined results from multiple collections, the groupings that were identified were not usable for addressing the research question. Instead of identifying groups based upon the content of the images, clustering instead identified the originating collection of the records. This is a fascinating result in that it demonstrates that the differing digitisation , metadata creation processes and terminologies used in different collections have a measurable effect on the resulting collections and that will effect search results. It does mean that the potential use clustering for co-reference identification in GLAM collections is likely to be limited.

In conclusion, this thesis demonstrates that a semi-automatic approach to co-reference identification is possible for photo-history records. In a search space spanning multiple collections this research shows that automatically locating promising co-reference matches is possible, but that automatically making the final determination of actual co-reference is not. The approach described locates potential matches that are of equivalent quality to those found by a manual approach and has a greater success rate at locating potential matches than that achieved by manually searching. As the majority of the search is conducted automatically, human intervention is only required for the final stage. This new

approach to searching GLAM collections means that photo-historians and other searchers are able to conduct more searches and search more widely than would otherwise be possible.

This thesis has successfully shown that it is possible to use CI techniques to assist in searching GLAM collection records and so reduce the difficulties experienced by current GLAM collection users.

8.0.1 Main contributions

A novel method for ordering potentially co-referent records from photographic collections which significantly reduces the number of actions required by the searcher whilst still maintaining equivalent result quality. The reduction in search actions is expected to significantly reduced person hour requirements when compared to manually searching.

A new, computationally efficient method for calculating semantic similarity between short text sections. Testing demonstrates a statistically insignificant drop in result quality but a significant decrease in computational time.

A new method for comparing person entity names which can handle unordered name elements in a computationally efficient manner.

A new method for comparing photographic processes which acknowledges and can handle the problem of process misidentification.

The exclusion of clustering as a viable solution to the research question, or at least a demonstration of problems with clustering and the identification of the cause of those issues.

8.1 With the benefit of hindsight

At the conclusion of this project there are certain actions which, although they appeared reasonable at the time, would not have been done or would have been done differently with the benefit of hindsight. Most significant of these regrets is however the engagement with members of the GLAM community. Is this research could be

restarted a much greater engagement with the GLAM community would have been maintained. This would have affected the research in two main ways. Firstly, the initial mindset going into this problem was that of a typical record linkage problem, such as is often seen in commercial customer database systems. While it was apparent from an early stage that syntax independence of the collection metadata would be challenging, just how heterogeneous the records were was surprising. When attempting to produce a similarity metric for the *process* field for example, the initial approach was to try and co-opt an existing hierarchy. It became apparent however that there was nothing suitable already available. The only way to address the information contained in the records was to use the expertise of experts in the field as their many years of experience meant that they already and instinctively knew the many names, groupings and oddities of photographic processes. The research would have gone much faster if, instead of trying to locate and re-purpose existing techniques, a greater respect for the unusual nature of the search space had been had and the instinctual knowledge of domain experts had been consulted sooner.

Secondly, a greater breadth of testing. Partially the limited number of testing respondents was a result of the considerable amount of time that testing took[12]. However, if a greater involvement with the GLAM community had been had during the result of the project it seems likely that more participants could have been recruited, this would have resulted in a greater sample size and a more comprehensive sampling of the domain. As it is, interactions with the GLAM community were mostly limited to a few key individuals and this resulted in a limited ability to recruit participants testing the proposed approach.

8.2 Further work

Although the new search approach is effective at locating co-reference possibilities, there are some clear areas for further investigations/improvement.

Reducing the computational requirements. The amount of time required to generate and the memory requirement for the subsequent similarity matrices are the significant issues with the overall search approach. It would be possible to temporarily solve the memory requirements problem of the proposed approach by deploying better hardware with greater amounts of memory. Given, however, that

[12] 60+ minutes of searching in some cases, plus the time to complete the questionnaire (40+ minutes).

the memory requirements of a similarity matrix increase exponentially as the number of items being compared increases and that the proposed approach makes use of multiple matrices, this would likely only be a temporary solution. Therefore, if this research is to be taken further, the primary focus would have to be reducing these requirements.

One factor which was made clear during the user testing of the new approach[13] is that, regardless of the number of records which are returned from the external collections, only a few tens of records at the very top of the generated dendrograms are ever examined. The exact number of records examined varies per person and per search but in cases where thousands or tens of thousands of results were collected from the external collections only a small fraction of the full dendrograms which were produced were used. The time spent generating the remaining portions of the dendrogram was, in effect, wasted. Therefore, future work on this research would be well advised to investigate methods of generating dendrograms containing just the top n records. Methods of generating these top n dendrograms without generating full similarity matrices for the individual fields should also be closely looked at.

Proposed approach person hour savings. Whilst the proposed approach can be expected to produce time savings versus manual searching due to the amount of automation that the proposed approach provides, this thesis has not been able to demonstrate these savings quantitatively. This is due to the present state of the software written during this research which placed greater importance on flexibility and demonstrating result quality than it did on computational efficiency. Subject to the improvements discussed in the paragraphs above, an investigation into the precise efficiency gains of the proposed approach versus manual searching would be an important contribution.

Wider searching. During the testing of the search approach, six GLAM collections were accessed: the Brooklyn Museum (BkM), DigitalNZ (New Zealand) (DNZ), Exhibitions of the Royal Photographic Society (ERPS), the Library of Congress (LoC), Photographic Exhibitions in Britain (PEiB) and the Victoria and Albert Museum (V&A). This represents just a fraction of the GLAM collections available. Whilst it is unrealistic to think that every GLAM collection could be included given the quantity and general lack of API access, those collections

[13] See section 7.1.1 on page 132.

with APIs could be easily included[14]. However, the issues regarding the high computational requirements of the proposed search approach need to be addressed first.

Additional field similarity metrics. The proposed search approach uses four separate fields when comparing records, *title*, *person*, *process* and *date*. Whilst this constitutes the majority of the fields in the ERPS collection, other collections have additional fields which could be used and compared if similarity metrics were created for them. Examples of these include where the photograph was taken and the physical size of the photograph. The inclusion of additional fields would not be of use in identifying co-referent matches for the ERPS records, but could be of assistance if the search approach was used to locate matches for records from other collections.

For the particular GLAM records investigated in this research, the *description* field in the ERPS records would be the most interesting candidate for inclusion. Whilst it was not possible to create an effective similarity metric as part of this research, the *description* field contains valuable information which should not be ignored. Direct comparison between the *description* fields of multiple records is very difficult and would probably prove to be ineffective given the level of variation in this field between records. However, the *description* field could be a valuable resource in populating other empty fields in a record. Amongst the ERPS collection records, various *description* fields contain information on the photographs, where they were taken and the photographic processes used. In cases where a field in a record is unpopulated, the *description* field could analysed for person/process/place names which could then be used to fill in the gaps in the record. Information extraction systems/approaches already exist which could be used to achieve this[26, 179, 225].

Improving individual field similarity metrics. Due to the computational requirements of generating pair-wise similarity matrices, the field similarity metrics used to create those matrices needed to be computationally fast and, therefore, relatively simple. If the full matrix generation requirement was removed[15], this could allow for more computationally expensive field similarity metrics to be used. The *title* metric would be an obvious candidate for this with existing approaches

[14] I.e. the British Museum[55] or Culturegrid[4] APIs.
[15] As discussed previously.

(e.g. STASIS) with known improvements in the quality of the similarity values already available.

Result presentation. Whilst the internal processes of the new search approach may result in a dendrogram, there is no reason why that the results need to presented to the users as such. Although some searchers may wish to view the links and connections between the results, a more traditional view of the records (i.e. grids, lists) may be beneficial. If such an approach were used, care would need to be taken to ensure that the record ordering within the results still made sense under the new layout.

Bibliography

[1] Metacrap: Putting the torch to seven straw-men of the meta-utopia. URL http://www.well.com/~doctorow/metacrap.htm. Accessed October 2011.

[2] British museum semantic web collection online, . URL http://collection.britishmuseum.org/. Accessed June 2014.

[3] The CIDOC conceptual reference model, . URL http://www.cidoc-crm.org/. Accessed March 2014.

[4] Culturegrid, . URL http://www.culturegrid.org.uk/. Accessed March 2013.

[5] Dublin core metadata element set, version 1.1, . URL http://dublincore.org/documents/dces/. Accessed June 2013.

[6] Dcmi frequently asked questions (faq), . URL http://dublincore.org/resources/faq/. Accessed June 2013.

[7] Museum handbook, part iii: Museum collection use. Technical report, National Park Service, 1998.

[8] Status of technology and digitization in the nation's museums and libraries. Technical report, Institute of Museum and Library Services, Washington, DC, 2006.

[9] Chin guide to museum standards. Technical report, Canadian Heritage Information Network (CHIN), 2013.

[10] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *Special Interest Group on Management of Data (SIGMOD) Record*, 27(2):94–105, June 1998.

[11] J. Allan. Relevance feedback with too much data. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 337–343, New York, NY, USA, 1995. ACM.

[12] A. Amin, J. van Ossenbruggen, L. Hardman, and A. van Nispen. Understanding cultural heritage experts' information seeking needs. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 39–47, New York, NY, USA, 2008. ACM.

[13] K. S. R. Anjaneyulu. Expert systems: An introduction. *Resonance*, 3(3): 46–58, 1998.

[14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In Karl Aberer, editor, *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, Berlin, 2007.

[15] S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats — an extensible framework for high-performance dataset analytics. In *Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management*, EKAW'12, pages 353–362, Berlin, Heidelberg, 2012. Springer-Verlag.

[16] Marc B. 52 weeks of inspiring illustrations, week 40: The photographic postcard, March 2013. URL `http://standrewsrarebooks.wordpress.com/2013/03/26/52-weeks-of-inspiring-illustrations-week-40-the-photographic-postcard/`. Accessed July 2013.

[17] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM press, New York, NY, USA, 1999.

[18] J. T. Ball. Can NLP systems be a cognitive black box?(is cognitive science relevant to ai problems?). In *AAAI Spring Symposium: Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, pages 1–6, 2006.

[19] R. Baxter, P. Christen, and T. Churches. A comparison of fast blocking methods for record linkage. In *ACM SIGKDD '03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 25–27, 2003.

[20] H. Bay, T. Tuytelaars, and L. Gool. SURF: Speeded up robust features. In Ale Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer, Berlin, 2006.

[21] J. E. Beaudoin. *An investigation of image users across professions: A framework of their image needs, retrieval and use*. PhD thesis, Drexel University, 2009.

[22] R. Bellman, R. Kalaba, and L. A. Zadeh. Abstraction and pattern classification. *Journal of Mathematical Analysis and Applications*, 13(1):1 – 7, 1966.

[23] T. Berners-Lee. *Weaving the Web : the past, present and future of the World Wide Web by its inventor.* Texere, London, 2000.

[24] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[25] J. C. Bezdek and R. J. Hathaway. VAT: a tool for visual assessment of (cluster) tendency. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN)*, volume 3, pages 2225 –2230, 2002.

[26] D. M. Bikel, R. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, 1999.

[27] B. Billerbeck and J. Zobel. Questioning query expansion: an examination of behaviour and parameters. In *Proceedings of the 15th Australasian database conference*, volume 27 of *ADC '04*, pages 69–76, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[28] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly, Beijing, 2009.

[29] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3): 1–22, 2009.

[30] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.

[31] B. G. Buchanan and E. H. Shortliffe. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley Series in Artificial Intelligence).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984.

[32] M. K. Buckland and F. C. Gey. The relationship between recall and precision. *Journal of the American Society for Information Science (JASIS)*, 45(1):12–19, 1994.

[33] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic Query Expansion Using SMART: TREC 3. In *The third Text REtrieval Conference (TREC).* National Institute of Standards & Technology, 1994.

[34] L. M. Chan and M. L. Zeng. Metadata interoperability and standardization–a study of methodology part i. *D-Lib magazine*, 12(6):1082–9873, 2006.

[35] H. Chen and K. J. Lynch. Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics (SMC)*, 22(5):885 –902, 1992.

[36] A. Chevalier, P. Rozencwajg, and B. Desjours. Impact of prior knowledge and computer interface organization in information searching performances: A study comparing younger and older web users. In C. Stephanidis, editor, *HCI International 2011 Posters Extended Abstracts*, volume 173 of *Communications in Computer and Information Science*, pages 373–377. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-22097-5. doi: 10.1007/978-3-642-22098-2_75. URL `http://dx.doi.org/10.1007/978-3-642-22098-2_75`.

[37] P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, 2012.

[38] P. Christen and K. Goiser. Assessing deduplication and data linkage quality: What to measure? In *Proceedings of the 2005 Australian Conference on Data Mining*, 2005.

[39] E. Coburn, R. Light, G. McKenna, R. Stein, and A. Vitzthum. LIDO - lightweight information describing objects version 1.0. Technical report, ICOM, 2010.

[40] A. Cocchi. Europeana: Opening new links in digital cultural heritage - a semantic approach to aggregate meaning. Master's thesis, UNIVERSITy of BOLOGNA, 2011.

[41] W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Information Integration on the Web (IIWeb)*, pages 73–78, 2003.

[42] J. Cousins and E. Niggemann. Europeana think culture: Strategic plan (2011-2015). Technical report, Europeana Foundation, January 2011.

[43] K. Coyle. Understanding metadata and its purpose. *The Journal of Academic Librarianship*, 31(2):160–163, 2005.

[44] J. W. Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage, Thousand Oaks, CA, 2 edition, 2003.

[45] D. Croft. Improving record matching in imprecise and uncertain datasets. *Literary and Linguistic Computing*, 27(4):347–354, 2012.

[46] D. Croft, S. Brown, and S. Coupland. Improving record matching across disparate historical resources. In *Proceedings of the 2012 Digital Humanities Congress (DHC)*, September 2012.

[47] D. Croft, S. Brown, and S. Coupland. A hybrid approach to co-reference identification within museum collections. In *2013 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES)*, pages 110–117. IEEE, April 2013.

[48] D. Croft, S. Coupland, J. Shell, and S. Brown. A fast and efficient semantic short text similarity metric. In *2013 UK Workshop on Computational Intelligence (UKCI)*, 2013.

[49] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, March 1964.

[50] J. Dawson. Suffix removal for word conflation. *Bulletin of the Association for Literary & Linguistic Computing (ALLC)*, 2(3):33–46, 1974.

[51] S. C. Deerwester, S. T. Dumais, G. W. Furnas, R. A. Harshman, T. K. Landauer, K. E. Lochbaum, and L. A. Streeter. Computer information retrieval using latent semantic structure. Patent, 1989. US 4839853, filed 1988.

[52] S. C. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Hasrhman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science (JASIS)*, 41(6):391 – 407, 1990.

[53] J. E. Dennis. Nonlinear least-squares. *State of the Art in Numerical Analysis*, 1977.

[54] K. Devlin. *Sets, functions, and logic : an introduction to abstract mathematics.* Chapman & Hall, 1992.

[55] C. Doctorow. The british museum. URL `http://www.britishmuseum.org/`. Accessed March 2013.

[56] D. L. Driscoll, A. Appiah-Yeboah, P. Salib, and D. J Rupert. Merging qualitative and quantitative data in mixed methods research: How to and why not. *Ecological and Environmental Anthropology (University of Georgia)*, page 18, 2007.

[57] H. L. Dunn. Record linkage. *American Journal of Public Health*, 36(12): 1412–1416, December 1946.

[58] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics and Systems*, 3(3): 32–57, 1973.

[59] W3C Semantic Web Education and Outreach group. Sweoig/taskforces/communityprojects/linkingopendata - w3c wiki, September 2013. URL `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`. Accessed September 2013.

[60] K. El-Arini and C. Guestrin. Beyond keyword search: Discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 439–447, New York, NY, USA, 2011. ACM.

[61] G. Elliott and N. Johnson. All the right letters just not necessarily in the right order.spelling errors in a sample of GCSE english scripts. In *Proceeding of the British Educational Research Association (BERA) 2008 Annual Conference*. Heriot Watt University, September 2005.

[62] Europeana Foundation. Europeana. URL `http://www.europeana.eu/portal/`. Accessed March 2013.

[63] M. P. Fay and M. A. Proschan. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, 4:1–39, 2010.

[64] E. A. Feigenbaum. The art of artificial intelligence: Themes and case studies of knowledge engineering. Technical report, Stanford University, 1977.

[65] E. A. Feigenbaum and B. G. Buchanan. DENDRAL and META-DENDRAL: Roots of knowledge systems and expert system applications. *Artificial Intelligence*, 59(1):233–240, 1993.

[66] C. Fellbaum. Wordnet and wordnets. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford, 2005. Elsevier.

[67] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association (JASA)*, 64:1183–1210, 1969.

[68] R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[69] Office for National Statistics. Internet access - households and individuals, 2011, August 2011. URL `http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-226727`. Accessed September 2013.

[70] Wikimedia Foundation. Wikipedia. URL `http://www.wikipedia.org/`. Accessed September 2013.

[71] Wikimedia Foundation. Wikipedia statistics, July 2012. URL `http://stats.wikimedia.org/EN/Sitemap.htm`. Accessed July 2012.

[72] W. N. Francis and H. Kucera. Brown corpus manual: Manual of information to accompany a standard corpus of present-day edited american english for use with digital computers. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[73] H. Frigui and O. Nasraoui. Simultaneous clustering and dynamic keyword weighting for text documents. In M. Berry, editor, *Survey of Text Mining: Clustering, Classification, and Retrieval*, chapter 3, pages 45–72. Springer, 2003.

[74] E. Gabrilovich and S. Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34:443–498, March 2009.

[75] G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*. Society for Industrial & Applied Mathematics (SIAM), Philadelphia, Pa. Alexandria, Va, 2007.

[76] L. Gill. OX-LINK: The oxford medical record linkage system. In *Record linkage techniques - Proceedings of an International Workshop and Exposition*, pages 15–33, March 1997.

[77] M. Gordon and M. Kochen. Recall-precision trade-off: A derivation. *Journal of the American Society for Information Science (JASIS)*, 40(3):145–151, 1989.

[78] S. J. Grannis, J. M. Overhage, and C. McDonald. Real world performance of approximate string comparators for use in patient matching. In *MEDINFO 2004: Proceedings of the 11th World Congress on Medical Informatics*, pages 43–47, 2004.

[79] G. Gregory. Automatic thesaurus generation from raw text using knowledge-poor techniques. In *Making Sense of Words. 9th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research*, 1993.

[80] M. R. Grossman and G. V Cormack. The grossman-cormack glossary of technology-assisted review. *Federal Courts Law Review*, 7(1), 2013.

[81] Museums Computer Group. museums computer group. URL `http://museumscomputergroup.org.uk/`. Accessed April 2014.

[82] K. Guney and N. Sarikaya. Comparison of mamdani and sugeno fuzzy inference system models for resonant frequency calculation of rectangular microstrip antennas. *Progress In Electromagnetics Research (PIER) B*, 12:81–104, 2009.

[83] M. M. Hall, P. Goodale, P. Clough, and M. Stevenson. The PATHS system for exploring digital cultural heritage. In Clare Mills, Michael Pidd, and Esther Ward, editors, *Proceedings of the 2012 Digital Humanities Congress (DHC)*, 2014.

[84] R. W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[85] D. J. Hand and K. Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review (ISR)*, 69(3):385–398, 2001.

[86] D. Harman. How effective is suffixing. *Journal of the American Society for Information Science (JASIS)*, 42:7–15, 1991.

[87] D. Harman. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 1–10, New York, NY, USA, 1992. ACM.

[88] A. Hart. *Knowledge acquisition for expert systems.* McGraw-Hill, New York, 1992.

[89] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, pages 203–212, 1999.

[90] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, and P. Fränti. Improving k-means by outlier removal. In *Image Analysis*, pages 978–987. Springer, 2005.

[91] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545. Association for Computational Linguistics, 1992.

[92] D. Henry and E. Brown. Using an RDF data pipeline to implement cross-collection search. In *Proceedings of International Conference for Culture and Heritage Online Museums and the Web 2012*, April 2012.

[93] G. Hibberd. Metaphors for discovery: A survey of library interfaces, March 2014. URL `http://georginahibberd.wordpress.com/2014/03/25/metaphors-for-discovery-a-survey-of-library-interfaces/`. Accessed May 2014.

[94] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[95] W. Hong, J. Y. L. Thong, W. Wong, and K. Y. Tam. Determinants of user acceptance of digital libraries: an empirical examination of individual differences and system characteristics. *Journal of Management Information Systems (JMIS)*, 18(3):97–124, 2002.

[96] D. Hood. Caverphone: Phonetic matching algorithm. Technical report, University of Otago, New Zealand, 2002.

[97] D. A Hull. Stemming algorithms - a case study for detailed evaluation. *Journal of the American Society for Information Science (JASIS)*, 47:70–84, 1996.

[98] D. A. Hull and G. Grefenstette. A detailed analysis of english stemming algorithms. Technical report, Xerox Research and Technology, 1996.

[99] P. Jackson. *Introduction to expert systems.* Addison-Wesley, Wokingham, England Reading, Mass, 1990.

[100] V. Jalali and M. R. M. Borujerdi. The effect of using domain specific ontologies in query expansion in medical field. In *International Conference on Innovations in Information Technology (IIT), 2008*, pages 277–281, Dec 2008.

[101] J. R. Jang and C. Sun. *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1997.

[102] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association (JASA)*, 84(406):414–420, 1989.

[103] K. Järvelin, J. Kristensen, T. Niemi, E. Sormunen, and H. Keskustalo. A deductive data model for query expansion. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 235–243. ACM, 1996.

[104] J. R. Jenkins. Where angels fear to tread: The problems of keyword search in e-discovery. Technical report, FTI Technology, 2010.

[105] M. F. Jiang, S. S. Tseng, and C. M. Su. Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, 22(6):691 – 700, 2001.

[106] JISC and Consortium of Research Libraries. Digitisation in the UK: the case for a UK framework. Technical report, JISC, 2005.

[107] N. K. Kasabov. *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*. MIT Press, Cambridge, MA, USA, 1st edition, 1996.

[108] A. Kaur and A. Kaur. Comparison of fuzzy logic and neuro fuzzy algorithms for air conditioning system. *International journal of soft computing and engineering (IJSCE)*, 2(1):2231–2307, 2012.

[109] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 74–75. ACM, 2002.

[110] A. Kennedy. The open roget's project. electronic lexical knowledge base. URL `http://rogets.site.uottawa.ca/`. Accessed September 2014.

[111] J. Klavans, R. Stein, S. Chun, and R. D. Guerra. Computational linguistics in museums: Applications for cultural datasets. In *Proceedings of International Conference for Culture and Heritage Online Museums and the Web 2011*. Archimuse, April 2011.

[112] E. Kolman and M. Margaliot. Knowledge extraction from neural networks using the all-permutations fuzzy rule base: The LED display recognition problem. In Joan Cabestany, Alberto Prieto, and Francisco Sandoval, editors,

Computational Intelligence and Bioinspired Systems, volume 3512 of *Lecture Notes in Computer Science*, pages 1222–1229. Springer, Berlin, 2005.

[113] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 191–202, New York, NY, USA, 1993. ACM.

[114] S. Krug. *Don't Make Me Think! A Common Sense Approach to Web Usability.* Que, 2000.

[115] K. L. Kwok, L. Grunfeld, and D. D. Lewis. TREC-3 ad hoc, routing retrieval and thresholding experiments using pircs. In *The Third Text REtrieval Conference (TREC)*, pages 247–256, Gaithersburg, Maryland, USA, 1995.

[116] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch. *Handbook of latent semantic analysis.* Psychology Press, 2013.

[117] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *British Machine Vision Conference*, 2004.

[118] K. Leszczyski, P. Penczek, and W. Grochulski. Sugeno's fuzzy measure and fuzzy clustering. *Fuzzy Sets and Systems*, 15:147 – 158, 1985.

[119] I. Lev, B. MacCartney, C. D. Manning, and R. Levy. Solving logic puzzles: From robust processing to precise semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, pages 9–16. Association for Computational Linguistics, 2004.

[120] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[121] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research (JMLR)*, 5:361–397, 2004.

[122] Y. Li, Z. A. Bandar, and D. Mclean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 15(4):871–882, 2003.

[123] Y. Li, D. Mclean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18(8):1138–1150, 2006.

[124] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[125] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2):209 – 261, 1993.

[126] A. Lipatov, A. Goncharuk, I. Helfenbein, V. Shilo, and V. Lehelt. Automatic creation of non-english wordnet-like lexical databases. In *Lecture Notes in Computer Science, Papillon Workshop 2003*, 2003.

[127] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129 – 137, March 1982.

[128] I. Lourdi, C. Papatheodorou, and M. Doerr. Semantic integration of collection description. *D-Lib Magazine*, 15:1082–9873, 2009.

[129] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[130] D. G. Lowe. Object recognition from local scale-invariant features. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.*, volume 2, pages 1150–1157 vol.2, 1999.

[131] H. Lüngen, C. Kunze, L. Lemnitzer, and A. Storrer. Towards an integrated OWL model for domain-specific and general language wordnets. In *Proceedings of the 4th Global Wordnet Conference (GWC)*, pages 281–296, 2008.

[132] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

[133] B. Magnini and M. Speranza. Merging global and specialized linguistic ontologies. In *Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases*, volume 4348, 2002.

[134] R. Mandala, T. Tokunaga, and H. Tanaka. Complementing wordnet with roget's and corpus-based thesauri for information retrieval. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 94–101. Association for Computational Linguistics, 1999.

[135] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60, 1947.

[136] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[137] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[138] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics (SIAM)*, 11(2):431–441, 1963.

[139] K. K. Matusiak. Information seeking behavior in digital image collections: A cognitive approach. *The Journal of Academic Librarianship*, 32(5):479 – 488, 2006.

[140] M. McIlroy. Development of a spelling list. *IEEE Transactions on Communications*, 30(1):91–99, 1982.

[141] J. M. Mendel. Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, 83(3):345–377, 1995.

[142] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence*, volume 1 of *AAAI'06*, pages 775–780. AAAI Press, 2006.

[143] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. ernock. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 196–201. IEEE Signal Processing Society, 2011.

[144] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, November 1995.

[145] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244, 1990.

[146] R. Mitton. Spellchecking by computer. *Journal of the Simplified Spelling Society*, 20(1):4–11, 1996.

[147] J. D. Moore and C. L. Paris. Requirements for an expert system explanation facility. *Computational Intelligence*, 7(4):367–370, 1991.

[148] J. D. Moore and W. R. Swartout. Explanation in expert systemss: A survey. Technical report, DTIC Document, 1988.

[149] A. Motro. Sources of uncertainty, imprecision, and inconsistency in information systems. In *Uncertainty Management in Information Systems*, pages 9–34. 1996.

[150] I. Mrazova, F. Mraz, M. Petricek, and Z. Reitermanov. Phonetic search in foreign texts. In *Artificial Neural Networks in Engineering (ANNIE)*, November 2008. Presentation at the Artificial Neural Networks in Engineering (ANNIE) conference in St. Lous, Missouri.

[151] Museums and the machine-processable web. Museum APIs. URL http://museum-api.pbworks.com/w/page/21933420/Museum%C2%A0APIs. Accessed September 2013.

[152] R. Navigli. Word sense disambiguation: A survey. *ACM Compututing Surveys*, 41(2):10:1–10:69, February 2009.

[153] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, October 1959.

[154] G. Niemeyer. python-dateutil. Computer software, March 2011. URL http://labix.org/python-dateutil.

[155] G. W. Nurcahyo, S. M. Shamsuddin, R. A. Alias, and M. N. M. Sap. Selection of defuzzification method to obtain crisp value for representing uncertain data in a modified sweep algorithm. *Journal of Computer Science and Technology (JCST)*, 3(2):22–28, 2003.

[156] M. Odell and R. Russell. The soundex coding system. Patent, 1918. US 1261167, filed 1917.

[157] The Library of Congress. American memory. URL http://memory.loc.gov/ammem/index.html. Accessed July 2013.

[158] University of Derby. Roger taylor. URL http://www.derby.ac.uk/graduation/honoraries/roger-taylor/. Accessed April 2014.

[159] U.S. National Library of Medicine. Medical subject headings. http://www.nlm.nih.gov/mesh/. Accessed July 2012.

[160] National Institute of Standards and Technology (NIST). Reuters corpora (rcv1, rcv2, trc2), 2004. URL http://trec.nist.gov/data/reuters/reuters.html. Accessed May 2012.

[161] National Institute of Standards and Technology (NIST). Text retrieval conference (trec), June 2013. URL http://trec.nist.gov/. Accessed June 2013.

[162] H. G. Oliveira and P. Gomes. Towards the automatic creation of a wordnet from a term-based lexical network. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 10–18. Association for Computational Linguistics, 2010.

[163] J. Oomen, L. B. Baltussen, and M. van Erp. Sharing cultural heritage the linked open data way: Why you should sign up. In *Proceedings of International Conference for Culture and Heritage Online Museums and the Web 2012*, volume 7, April 2012.

[164] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean. A comparative study of two short text semantic similarity measures. In *Proceedings of the 2nd KES International conference on Agent and multi-agent systems: technologies and applications*, KES-AMSTA'08, pages 172–181, Berlin, Heidelberg, 2008. Springer-Verlag.

[165] C. D. Paice. Another stemmer. *Special Interest Group on Information Retrieval (SIGIR) Forum*, 24(3):56–61, November 1990.

[166] F. Patman and L. Shaefer. Is soundex good enough for you? on the hidden risks of soundex-based name searching. Technical report, Language Analysis Systems, Inc., Herndon, 2001.

[167] M. Patton. *Qualitative research and evaluation methods*. Sage Publications, Thousand Oaks, Calif, 2002.

[168] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning*, pages 727–734. Morgan Kaufmann, 2000.

[169] F. J. Pelletier. Metamathematics of fuzzy logic. *The Bulletin of Symbolic Logic*, 6(3):342–346, 2000.

[170] U. Pfeifer, T. Poersch, and N. Fuhr. Retrieval effectiveness of proper name search methods. *Information Processing & Management*, 32(6):667 – 679, 1996.

[171] L. Philips. Anthropomorphic software. URL `http://amorphics.com/metaphone3.html`. Accessed January 2013.

[172] L. Philips. The double metaphone search algorithm. *C/C++ users journal*, 18(6):38–43, 2000.

[173] R. Poll. Numeric: statistics for the digitisation of european cultural heritage. *Program: electronic library and information systems*, 44(2):122–131, 2010.

[174] N. Poole. The cost of digitising europes cultural heritage a report for the comité des sages of the european commission. Technical report, Collections Trust, 2010.

[175] M. F. Porter. An algorithm for suffix stripping. In Karen S. Jones and Peter Willett, editors, *Readings in information retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[176] K. J. Preacher. Calculation for the test of the difference between two independent correlation coefficients. Software, May 2002. URL `http://quantpsy.org`.

[177] M. Pritchard. British photographic history. URL `http://britishphotohistory.ning.com/`. Accessed April 2014.

[178] J. Purday. Breaking new ground: Europeana annual report and accounts 2011. Technical report, Europeana Foundation, June 2012.

[179] Y. Ravin and N. Wacholder. IBM research report. extracting names from natural-language text. Technical report, IBM, 1997.

[180] J. M. Reilly. *Care and identification of 19th-century photographic prints*. Eastman Kodak Company, Rochester, NY, 1986.

[181] Microsoft Research. Microsoft research paraphrase corpus, March 2012. URL `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/`. Accessed July 2013.

[182] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[183] I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[184] A. Roberts. *Crash Course in Library Gift Programs: The Reluctant Curator's Guide to Caring for Archives, Books, and Artifacts in a Library Setting*. Libraries Unlimited, 2007.

[185] A. M. Robertson and P. Willett. A comparison of spelling-correction methods for the identification of word forms in historical text databases. *Literary and Linguistic Computing*, 8(3):143–152, January 1993.

[186] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 42–49, New York, NY, USA, 2004. ACM.

[187] J. Rubin. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[188] E. H. Ruspini. A new approach to clustering. *Information and control*, 15(1): 22–32, 1969.

[189] M. Sanderson and W. B. Croft. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451, 2012.

[190] M. C. Schraefel. Building knowledge: What's beyond keyword search? *Computer*, 42(3):52–59, 2009.

[191] C. Schwarz. Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, 41(6):408–417, 1990.

[192] P. P. Senellart and V. D. Blondel. Automatic discovery of similar words. In Michael Berry, editor, *Survey of Text Mining: Clustering, Classification, and Retrieval*, chapter 3, pages 25–43. Springer, 2003.

[193] S. C. Shapiro. *Encyclopedia of Artificial Intelligence*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition, 1992.

[194] S. Sharoff. Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT, 2006.

[195] A. Shiri and K. Molberg. Interfaces to knowledge organization systems in canadian digital library collections. *Online Information Review*, 29(6):604–620, 2005.

[196] A. F. Smeaton. Progress in the application of natural language processing to information retrieval tasks. *The computer journal*, 35(3):268–278, June 1992.

[197] C. Soanes and A. Stevenson, editors. *Oxford Dictionary of English, Second Edition, Revised*. Oxford University Press, 2006.

[198] S. Sofia, A. Ntoulas, K. Maria, and C. Dimitris. Expanding ewn with domain-specific terminology using common lexical resources: Vocabulary completeness and coverage issues. In *1st International Global WordNet Conference*, page 41. Central Institute of Indian Languages, 2002.

[199] Archer Software. Nominex, british surname matching system. URL `http://www.archersoftware.co.uk/nominex/`. Accessed June 2014.

[200] N. Srinivasan and J. Agrawal. The relationship between prior knowledge and external search. *Advances in consumer research*, 15(1):27–31, 1988.

[201] Lærd statistics. Mann-whitney u test using spss. URL `https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php`. Accessed May 2013.

[202] R Stein and E. Coburn. CDWA Lite and museumdat: New developments in metadata standards for cultural heritage information. *In Proceedings of the 2008 Annual Conference of CIDOC*, pages 15–18, 2008.

[203] A. Stow. Digitisation of museum collections: a worthwhile effort? Master's thesis, Institutionen för kulturvård, Göteborgs universitet, June 2011.

[204] N. Stroeker and R. Vogels. ENUMERATE survey report on digitisation in european cultural heritage institutions 2012. Technical report, ENUMERATE Thematic Network, May 2012.

[205] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[206] D. C. Stulik and A. Kaplan. A new scientific methodology for provenancing and authentication of 20th century photographs: nondestructive approach. In *9th International Conference on NDT of Art: Non-Destructive Investigations and Microanalysis for the Dianostics and Conservation of Cultural and Environmental Heritage*, May 2008.

[207] D. C. Stulik and A. Kaplan. Collaborative research: Working with the alternative photographic processes community. *The Getty Conservation Institute Newsletter*, 27:14–15, June 2012.

[208] D. C. Stulik and A. Kaplan. *The atlas of analytical signatures of photographic processes*, volume Collotype. The Getty Conservation Institute, Los Angeles, CA, 2013.

[209] M. Sugeno and G. T. Kang. Structure identification of fuzzy model. *Fuzzy sets and systems*, 28(1):15–33, 1988.

[210] R. L. Taft. *Name search techniques*. Number 1. Bureau of Systems Development, 1970.

[211] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics (SMC)*, SMC-15(1):116 –132, January 1985.

[212] Abbas Tashakkori and Charles Teddlie. *Sage handbook of mixed methods in social & behavioral research*. Sage, Thousand Oaks, CA, 2010.

[213] K. Thangavel and A. Pethalakshmi. Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing*, 9(1):1–12, 2009.

[214] P. Thompson, H. R. Turtle, B. Yang, and J. Flood. Trec-3 ad hoc retrieval and routing experiments using the win system. In *The third Text REtrieval Conference (TREC)*, pages 211–218, Gaithersburg, Maryland, USA, 1995.

[215] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[216] G. G. Towell and J. W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13(1):71–101, 1993.

[217] D. Tudhope, C. Binding, and K. May. Semantic interoperability issues from a case study in archaeology. In *Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop SIEDL*, pages 88–99, 2008.

[218] P. D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK, 2001. Springer-Verlag.

[219] De Montfort University. De montfort university. http://dmu.ac.uk/, . Accessed September 2013.

[220] De Montfort University. Exhibitions of the royal photographic society 1870-1915, . URL `http://erps.dmu.ac.uk/`. Accessed September 2013.

[221] De Montfort University. Photographic exhibitions in britain 1839-1865, . URL `http://peib.dmu.ac.uk/`. Accessed September 2013.

[222] De Montfort University. Photographic history research centre, . URL `http://www.dmu.ac.uk/research/research-faculties-and-institutes/art-design-humanities/phrc/photographic-history-research-centre-phrc.aspx`. Accessed September 2013.

[223] Princeton University. Wnstats(7wn) manual page, 2010. URL `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`. Accessed September 2013.

[224] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. Technical report, MICC, Maastricht University, 2009.

[225] M. van Erp, J. Oomen, R. Segers, C. van de Akker, L. Aroyo, G. Jacobs, S. Legne, L. van der Meij, J. R. van Ossenbruggen, and G. Schreiber. Automatic heritage metadata enrichment with historic events. In *Proceedings of International Conference for Culture and Heritage Online Museums and the Web 2011*. Archimuse, April 2011.

[226] L. Vanderwende, G. Kacmarcik, H. Suzuki, and A. Menezes. Mindnet: an automatically-created lexical resource. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 8–9. Association for Computational Linguistics, 2005.

[227] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 171–180, New York, NY, USA, 1993. ACM.

[228] T. M. Vu, P. O. Siebers, and C. Wagner. Comparison of crisp systems and fuzzy systems in agent-based simulation: A case study of soccer penalties. In *13th UK Workshop on Computational Intelligence (UKCI)*, pages 54–61, September 2013.

[229] C. Wagner. Juzzy–a java based toolkit for type-2 fuzzy logic. In *2013 IEEE Symposium on Advances in Type-2 Fuzzy Logic Systems (T2FUZZ)*. IEEE, 2013.

[230] S. Walker, S. E. Robertson, M. Boughanem, G. J. F. Jones, K. S. Jones, and P. Willett. Okapi at TREC-6 - automatic ad hoc, VLC, routing, filtering and QSDR. In *The Sixth Text REtrieval Conference (TREC)*, pages 125–136, January 1998.

[231] C. Warwick, M. Terras, P. Huntington, and N. Pappa. If you build it will they come? the LAIRAHstudy: quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. *Literary and Linguistic Computing*, 23(1):85–102, 2008.

[232] P. Wiemer-Hastings. Adding syntactic information to lsa. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 989–993. Morgan Kaufmann, 2000.

[233] B. M. Wildemuth. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(3):246–258, 2004.

[234] D. R. Wilson. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In *The 2011 International Joint Conference on Neural Networks (IJCNN)*, pages 9–14, July 2011.

[235] M. L. Wilson, B. Kules, M. C. Schraefel, and B. Shneiderman. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010.

[236] M. Wilz. Aspekte der kodierung phonetischer ahnlichkeiten in deutschen eigennamen. Master's thesis, Department of Linguistics, University of Cologne, 2005.

[237] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359, 1990.

[238] W. E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau, 1999.

[239] W. E. Winkler. Machine learning, information retrieval and record linkage. In *Proceedings of the Survey Research Methods Section American Statistical Association*, pages 20–29, 2000.

[240] K. Yang, R. Steele, and A. Lo. An ontology for xml schema to ontology mapping representation. 2007.

[241] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[242] L. A. Zadeh. Fuzzy logic. *Computer*, 21(4):83–93, 1988.

[243] L. A. Zadeh. Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, 37(3):77–84, 1994.

[244] M. L. Zeng and L. M. Chan. Metadata interoperability and standardization-a study of methodology, part ii. *D-Lib Magazine*, 12(6):1082–9873, 2006.

[245] J. Zobel and P. Dart. Phonetic string matching: Lessons from information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 166–172. ACM, 1996.

A
Collection records

A.1 BkM

Documentation for the BkM API can be found at `http://www.brooklynmuseum.org/opencollection/api/` Using the LoC REST API requires the use of carefully formatted Uniform Resource Locators (URLs) in order to return JSON formatted results. During this research the following represents the basic URL which was used for all queries run against the BkM collection, the actual URLs used would of course have been slightly modified depending on the query terms and the number of records returned by the query.

```
http://www.brooklynmuseum.org/opencollection/api/?method=
collection.search&version=1&item_type=object&format=json&collection_
id=3&api_key=[ApiUserKey]&results_limit=20&start_index=
[GroupOfRecordsToReturn]&keyword=[SearchTerms]
```

A.1.1 Example data

The BkM record fields and the ERPS fields used in the individual similarity metrics were paired up as shown below.

- *Title* - title
- *Description* - inscribed (not present in example record)
- *Person* - artists → name
- *Process* - medium
- *Date* - object_date_begin and object_date_end combined.

Included below is a single record taken from the JSON file returned for the search term “apple”. A single record is, however, all that is required to demonstrate the record format used by the BkM.

```
1 {
2   "type": "object",
```

```
  "id": "124064",
  "title": "Snake Skeleton with Apple",
  "uri": "http:\/\/www.brooklynmuseum.org\/opencollection
    \/objects\/124064\/Snake_Skeleton_with_Apple",
  "images":
  {
    "total": "1",
    "results_limit": 1,
    "0":
    {
      "uri": "http:\/\/cdn2.brooklynmuseum.org\/images\/
        opencollection\/objects\/size0\/1989.190.1_PS2.
        jpg",
      "thumb_uri": "http:\/\/cdn2.brooklynmuseum.org\/
        images\/opencollection\/objects\/size0\/1989.190
        .1_PS2.jpg",
      "credit": "Brooklyn Museum photograph",
      "description": null,
      "is_color": true,
      "rank": 0
    }
  },
  "accession_number": "1989.190.1",
  "object_date": "Early 1950s",
  "object_date_begin": "1950",
  "object_date_end": "1954",
  "medium": "Vintage gelatin silver photograph",
  "dimensions": "9 1\/2 x 8 3\/8 in.",
  "credit_line": "Gift of Eileen and Adam Boxer",
  "classification": "Photograph",
  "artists": [
    {
      "uri": "http:\/\/www.brooklynmuseum.org\/
        opencollection\/artists\/7999\/Pierre_Jahan",
      "id": "7999",
      "name": "Pierre Jahan",
      "dates": null,
      "nationality": null,
      "role": "Artist",
      "type": "artist"
    } ],
  "collection": "Photography",
  "rightstype": "copyright_artist_or_artists_estate",
  "rank": 0
```

```
}
```

A.2 DNZ

Documentation for the DNZ API can be found at `http://www.digitalnz.org/developers`. Using the LoC REST API requires the use of carefully formatted URLs in order to return JSON formatted results. During this research the following represents the basic URL which was used for all queries run against the DNZ collection, the actual URLs used would of course have been slightly modified depending on the query terms and the number of records returned by the query.

```
http://api.digitalnz.org/records/v1.json?&api_key=[ApiUserKey]
&search_text=category:Images+'[SearchTerms]'&num_results=100&start=
[GroupOfRecordsToReturn]
```

A.2.1 Example data

The DNZ record fields and the ERPS fields used in the individual similarity metrics were paired up as shown below.

- *Title* - title
- *Description* - description
- *Person* - author
- *Date* - display_date

Included below is a single record taken from the JSON file returned for the search term "apple". A single record is, however, all that is required to demonstrate the record format used by the DNZ.

```
{
"id": 30618963,
"title": "Loading Apple Cases, 1958",
"alternate_title": null,
"description": "",
"additional_description": null,
"content_provider": "Kete Tasman",
"display_content_partner": "Kete Tasman",
"collection_title": "Kete Tasman",
"display_collection": "Kete Tasman",
"primary_collection": "Kete Tasman",
"contributing_partner": null,
"category": "Images",
```

```
"author": "Tasman District Libraries Kete",
"contributor": null,
"object_copyright": "All rights reserved",
"citation": null,
"credit_creator": "",
"language": null,
"provenance": "",
"publisher": "ketetasman.peoplesnetworknz.info",
"rights": "http://ketetasman.peoplesnetworknz.info/about/topics/show/4-terms-and-conditions",
"usage": "All rights reserved",
"source": null,
"tag": null,
"thesis_level": "",
"holding": null,
"library_collection": null,
"shelf_location": "",
"eprints_type": null,
"text": "",
"fulltext": "",
"dctype": null,
"dnz_type": "Unknown",
"format": "image/pjpeg",
"dc_identifier": null,
"date": null,
"display_date": "",
"published_date": null,
"syndication_date": "2013-03-26T05:55:55+13:00",
"display_url": "http://ketetasman.peoplesnetworknz.info/site/images/show/545-loading-apple-cases-1958",
"large_thumbnail_url": "http://ketetasman.peoplesnetworknz.info/image_files/0000/0000/2723/Loading-Apple-cases-1958_large.jpg",
"object_rights_url": "http://ketetasman.peoplesnetworknz.info/about/topics/show/4-terms-and-conditions",
"thumbnail_url": "http://ketetasman.peoplesnetworknz.info/image_files/0000/0000/2723/Loading-Apple-cases-1958_medium.jpg",
"origin_url": "",
"metadata_url": "",
"object_url": "http://ketetasman.peoplesnetworknz.info/image_files/0000/0000/2723/Loading-Apple-cases-1958.jpg",
"atl_free_download": "",
```

```
"atl_physical_viewability": "",
"atl_purchasable": "",
"atl_purchasable_download": "",
"atl_location_code": "",
"atl_usage_code": "",
"anzsrc_code": "",
"marsden_code": null,
"subject": "Riverside Community",
"coverage": null,
"source_url": "http://api.digitalnz.org/records/30618963/
   source",
"geo_co_ords": ",",
"collection_parent": null,
"collection_root": null
},
```

A.3 ERPS

The ERPS can be viewed online at `http://erps.dmu.ac.uk/`. As it lacks a REST API a copy of the underlying database was used instead. This was possible as the ERPS collection is hosted by DMU. Although the full set of ERPS collection records was available, only those records with image data were actually included.

A.4 LoC

Documentation for the LoC API can be found at `http://www.loc.gov/pictures/api`. Using the LoC REST API requires the use of carefully formatted URLs in order to return JSON formatted results. During this research the following represents the basic URL which was used for all queries run against the LoC collection, the actual URLs used would of course have been slightly modified depending on the query terms and the number of records returned by the query.

`http://www.loc.gov/pictures/search/?fa=displayed%3Aanywhere&fo=json&c=100&q=[SearchTerms]&sp=[GroupOfRecordsToReturn]`

A.4.1 Example data

The LoC record fields and the ERPS fields used in the individual similarity metrics were paired up as shown below.

- *Title* - title
- *Description* - medium_brief

- *Person* - creator
- *Process* - medium
- *Date* - created_published_date

Included below is a single record taken from the JSON file returned for the search term "apple". A single record is, however, all that is required to demonstrate the record format used by the LoC.


```
{
  "source_created": "1993-01-28 00:00:00",
  "index": 5,
  "medium": "2 photographic prints on album page : silver
     gelatin.",
  "reproduction_number": "LC-USZC2-4155 (color film copy
    slide)",
  "links":
  {
    "item": "http://www.loc.gov/pictures/item/gsc19940282
      67/PP/",
    "resource": "http://www.loc.gov/pictures/item/gsc1994
      028267/PP/resource/"
  },
  "title": "Seventy-one years, or, My life with
    photography. Kitchen of the John Howard Payne house,
     June 24, 1924; Windmill and apple tree, May 23, 192
    4",
  "image":
  {
    "alt": "digitized item thumbnail",
    "full": "http://www.loc.gov/pictures/lcweb2/service/
      pnp/gsc/5a00000/5a00000/5a00048r.jpg",
    "square": "http://lcweb2.loc.gov/service/pnp/gsc/5a00
      000/5a00000/5a00048_75x75px.jpg",
    "thumb": "http://lcweb2.loc.gov/service/pnp/gsc/5a000
      00/5a00000/5a00048_150px.jpg"
  },
  "created": "2013-11-27 00:00:00",
  "modified": "2013-11-27 00:00:00",
  "collection": [ "ammem", "diof", "gsc", "pp" ],
  "creator": "Gottscho, Samuel H. (Samuel Herman), 1875-1
    971",
  "call_number": "LOT 12400, p. 022 <P&P> [P&P]",
  "medium_brief": "2 photographic prints on album page :"
    ,
```

```
25    "source_modified": "2010-11-26 00:00:00",
26    "pk": "gsc1994028267/PP",
27    "created_published_date": "photographed 1924, printed
        later.",
28    "subjects": [ "Dwellings.", "Windmills.", "United
        States--New York (State).", "Silver gelatin prints."
         ]
29  }
```

A.5 PEiB

The PEiB can be viewed online at `http://peib.dmu.ac.uk/`. As it lacks a REST API a copy of the underlying database was used instead. This was possible as the PEiB collection is hosted by DMU. The full set of PEiB collection records were included.

A.6 V&A

Documentation for the V&A API can be found at `http://www.vam.ac.uk/api/`. Using the V&A REST API requires the use of carefully formatted URLs in order to return JSON formatted results. During this research the following represents the basic URL which was used for all queries run against the V&A collection, the actual URLs used would of course have been slightly modified depending on the query terms and the number of records returned by the query.

`http://www.vam.ac.uk/api/json/museumobject/search?objectnamesearch=photograph&q='[SearchTerms]'&limit=100&offset=[GroupOfRecordsToReturn]`

The V&A API has an additional complication not present in the other REST interfaces used during this research, if the number of results exceeds 2,000 then only the first 2,000 are returned. The specific focus of this research is finding matches for the ERPS records and those were all exhibited between 1870 and 1915. Therefore, when more than 2,000 results were found the URL was modified to include a further two parameters which returned only those records from between 1860 and 1925. Both the date filtered and un-filtered URLs were then used to query the collection record. The un-filtered URL would collect 2,000 records and the filtered URL would collect some number of additional records but ones which were expected to have a higher chance of matching the ERPS records.

In the event of the 2,000 result limit being reached the URL would follow the following format.

`http://www.vam.ac.uk/api/json/museumobject/search?objectnamesearch=photograph&q='[SearchTerms]'&limit=100&offset=`

`[GroupOfRecordsToReturn]&after=1860&before=1925`

Not all the required record information is returned in the initial JSON files. Under the V&A API detailed record information is retrieved on a record by record basis. In order to collect all of the information required therefore, another URL must be created and queried for each individual record. These URLs follow the format shown below, where [objectnumber] corresponds to the values taken from the object_number fields in the initial JSON files.

`http://www.vam.ac.uk/api/json/museumobject/[objectnumber]`

A.6.1 Example data

The V&A record fields in the detailed record files and the ERPS fields used in the individual similarity metrics were paired up as shown below.

- *Title* - title
- *Description* - physical_description
- *Person* - artist
- *Process* - materials_techniques
- *Date* - date_start

Included below is a single record taken from the JSON file returned for the search term "apple" and the in-depth record information file for that same record.


```
{
  "pk": 36281,
  "model": "collection.museumobject",
  "fields":
  {
    "primary_image_id": "",
    "rights": 2,
    "year_start": 1999,
    "object_number": "O41240",
    "artist": "Jones, Sarah",
    "museum_number": "E.801-2000",
    "object": "Photograph",
    "longitude": null,
    "last_processed": "2014-01-31 20:25:11",
    "event_text": "",
    "place": "",
    "location": "In Storage",
    "last_checked": "2014-01-31 20:25:11",
    "museum_number_token": "e8012000",
```

```
    "latitude": null,
    "title": "The ^Apple Tree, Charlton I",
    "date_text": "1999 (made)",
    "slug": "the-apple-tree-charlton-i-photograph-jones-sarah",
    "sys_updated": "2013-08-25 00:00:00",
    "collection_code": "PDP"
  }
}
```

```
[
{
"pk": 36281,
"model": "collection.museumobject",
"fields": {
"original_price": "",
"attributions_note": "",
"related_museum_numbers": "",
"museum_number": "E.801-2000",
"date_end": "1999-12-31",
"labels": [
{
"pk": 6905,
"model": "collection.label",
"fields": {
"date": "",
"museumobject": 36281,
"label_text": "Sarah Jones (born London 1959)\nApple Tree (Charlton) II\n1999\nC-type print\n\nSarah Jones is among the leading contemporary artists that are making carefully staged, large-scale colour photographs. The proportions of her photographs accentuate the relationship between the almost life-size subject and the viewer. \n\nThis still and enigmatic scene contains gestures and objects that suggest psychological depth and meaning. A girl stands in a suburban garden in front of an apple tree. She holds a frog or toad. The juxtaposition of the pattern on her tee-shirt with the fruit and flowers that surround her sets up a correspondence between the urban and the natural. Strongly lit against the dark background, these elements gain a symbolic resonance but also seem eerily artificial."
}
```

```
}
],
"descriptive_line": "Photograph, 'The Apple Tree,
  Charlton I', by Sarah Jones, 1999",
"shape": "",
"longitude": null,
"year_start": 1999,
"exhibitions": [
{
"pk": 1347,
"model": "collection.exhibition",
"fields": {
"va": true,
"venue_id": 3747,
"year_start": 2004,
"name": "History of Photography",
"date_end": "2004-10-07",
"museumobject_count": 5,
"venue": "Photography gallery",
"date_start": "2004-10-07",
"year_end": 2004,
"source": "",
"cis_id": null,
"museumobject_image_count": 4,
"type": "",
"slug": "history-of-photography",
"date_text": "07/10/2004"
}
}
],
"subjects": [
{
"pk": 47877,
"model": "collection.subject",
"fields": {
"name": "girl",
"museumobject_count": 532,
"source": "",
"cis_id": "x47814",
"museumobject_image_count": 311,
"type": "",
"slug": "girl"
}
},
```

```
{
"pk": 24306,
"model": "collection.subject",
"fields": {
"name": "garden",
"museumobject_count": 360,
"source": "object",
"cis_id": "24993",
"museumobject_image_count": 201,
"type": "",
"slug": "garden"
}
},
{
"pk": 957,
"model": "collection.subject",
"fields": {
"name": "apple tree",
"museumobject_count": 14,
"source": "object",
"cis_id": "x30173",
"museumobject_image_count": 8,
"type": "",
"slug": "apple-tree"
}
}
],
"date_text": "1999 (made)",
"primary_image_id": "",
"rights": 2,
"physical_description": "Photograph depicting a girl
  standing in a suburban garden in front of an apple
  tree. She holds a frog or toad.",
"dimensions": "Height: 150 cm, Width: 150 cm",
"title": "The ^Apple Tree, Charlton I",
"date_start": "1999-01-01",
"materials_techniques": "C-type print",
"last_processed": "2014-01-31 20:25:11",
"label": "Sarah Jones (born London 1959)\nApple Tree (
  Charlton) II\n1999\nC-type print\n\nSarah Jones is
  among the leading contemporary artists that are making
   carefully staged, large-scale colour photographs. The
   proportions of her photographs accentuate the
  relationship between the almost life-size subject and
```

```
    the viewer. \n\nThis still and enigmatic scene
    contains gestures and objects that suggest
    psychological depth and meaning. A girl stands in a
    suburban garden in front of an apple tree. She holds a
     frog or toad. The juxtaposition of the pattern on her
     tee-shirt with the fruit and flowers that surround
    her sets up a correspondence between the urban and the
     natural. Strongly lit against the dark background,
    these elements gain a symbolic resonance but also seem
     eerily artificial.",
"event_text": "",
"production_type": "",
"collections": [
{
"pk": 1,
"model": "collection.collection",
"fields": {
"code": "PDP",
"name": "Prints, Drawings and Paintings Collection",
"museumobject_count": 716971,
"source": "",
"cis_id": null,
"museumobject_image_count": 135857,
"type": "",
"slug": "pdp"
}
}
],
"location": "In Storage",
"marks": "",
"latitude": null,
"techniques": [
{
"pk": 1072,
"model": "collection.technique",
"fields": {
"name": "C-type",
"museumobject_count": 18,
"source": "",
"cis_id": null,
"museumobject_image_count": 4,
"type": "",
"slug": "c-type"
}
```

```
}
],
"materials": [],
"edition_number": "",
"styles": [],
"inventory_set": [
{
"pk": 76957,
"model": "collection.inventory",
"extras": {
"gallery_id": null
},
"fields": {
"box": "",
"case": "",
"inventory_number": 662963,
"room": "",
"part_name": "",
"museum_number": "E.801-2000",
"museumobject": 36281,
"shelf": "",
"site": "",
"on_display": false,
"status": "",
"location": "In Storage",
"museum_number_token": "e8012000",
"gallery": null
}
}
],
"updated": null,
"galleries": [],
"names": [
{
"pk": 4558,
"model": "collection.name",
"fields": {
"death_date": null,
"surname": "",
"name": "Jones, Sarah",
"gender": null,
"museumobject_count": 4,
"death_year": null,
"source": "object_production",
```

```
"cis_id": "A4022",
"museumobject_image_count": 1,
"forename": "",
"birth_date": null,
"nationality": "",
"type": "person",
"slug": "jones-sarah",
"birth_year": null
}
}
],
"placecontext_set": [],
"original_currency": "",
"museum_number_token": "e8012000",
"object": "Photograph",
"categories": [
{
"pk": 45,
"model": "collection.category",
"fields": {
"name": "Photographs",
"museumobject_count": 40494,
"source": "cis_category",
"cis_id": null,
"museumobject_image_count": 31429,
"type": "",
"slug": "photographs"
}
}
],
"last_checked": "2014-01-31 20:25:11",
"public_access_description": "Sarah Jones is among the
  leading contemporary artists who are making carefully
  staged, large-scale colour photographs. The
  proportions of her photographs accentuate the
  relationship between the almost life-size subject and
  the viewer. \n\nThis still and enigmatic scene
  contains gestures and objects that suggest
  psychological depth and meaning. A girl stands in a
  suburban garden in front of an apple tree. She holds a
   frog or toad. The juxtaposition of the pattern on her
   tee-shirt with the fruit and flowers that surround
  her sets up a correspondence between the urban and the
   natural. Strongly lit against the dark background,
```

```
    these elements gain a symbolic resonance but also seem
     eerily artificial.",
"exhibition_history": "History of Photography (Victoria
    and Albert Museum 08/01/2003-30/04/2004)",
"bibliography": "",
"vanda_exhibition_history": "",
"slug": "the-apple-tree-charlton-i-photograph-jones-sarah
    ",
"sys_updated": "2013-08-25 00:00:00",
"image_set": [],
"places": [],
"artist": "Jones, Sarah",
"namecontext_set": [
{
"pk": 37503,
"model": "collection.namecontext",
"extras": {
"name_id": 4558
},
"fields": {
"name": 4558,
"part_name": "",
"uncertainty": "",
"museumobject": 36281,
"role": "artist",
"order": 1
}
}
],
"historical_significance": "",
"year_end": 1999,
"object_number": "O41240",
"events": [],
"credit": "Given by BMW Financial Services Group",
"history_note": "This photograph was presented to the
    museum by BMW Financial Services Group in return for
    professional advice in 1998-1999. Part of a entitled <
    i>Making Your Dreams Come True</i>  the work results
    from a commission awarded by BMW Financial Services
    Group.",
"place": "",
"production_note": "",
"historical_context_note": "",
"collection_code": "PDP"
```

```
245 }
246 }
247 ]
```

B
Title field

Algorithm 2 Algorithm for the *title* similarity metric.

Input: Two tokenised and stemmed text strings A and B
Output: Floating point value between 0.0 and 1.0

$W \leftarrow \{A + B\}$ ▷ Get a list of the distinct words in the two strings
$VA \leftarrow$ [A.count(w) for w in W] ▷ Create a term vector of string A
$VB \leftarrow$ [B.count(w) for w in W] ▷ Create a term vector of string B

for i in $|W|$ **do** ▷ Create the weighted vectors for both strings
  for j in $|W|$ **do**
    $WA_i \leftarrow WA_i + (VA_j \cdot \text{TermSim}(W_i, W_j))$
    $WB_i \leftarrow WB_i + (VB_j \cdot \text{TermSim}(W_i, W_j))$
  end for
end for

for i in $|W|$ **do** ▷ Calculate the cosine similarity of the weighted vectors
  $d \leftarrow d + (WA_i \cdot WB_i)$
  $da \leftarrow da + {WA_i}^2$
  $db \leftarrow db + {WB_i}^2$
end for

return $d/(\sqrt{da} \cdot \sqrt{db})$

procedure TermSim(A,B) ▷ Find the highest similarity between two words
  $s = 0$
  for a in Synsets(A) **do** ▷ Synsets gets the list of associated synsets from WordNet
    for b in Synsets(B) **do**
      $t \leftarrow \text{SynsetSimilarity}(a, b)$ ▷ Path distance, Wu-Palmar was also tried
      if $t > s$ **then**
        $s \leftarrow t$
      end if
    end for
  end for

  return s
end procedure

Sentence pair			Semantic similarity measure			
Id	A	B	Human	*title*	STASIS	LSA
1	cord	smile	0.010	0.180	0.329	0.510
5	autograph	shore	0.005	0.198	0.287	0.530
9	asylum	fruit	0.005	0.280	0.209	0.505
13	boy	rooster	0.108	0.166	0.530	0.535
17	coast	forest	0.063	0.324	0.356	0.575
21	boy	sage	0.043	0.324	0.512	0.530
25	forest	graveyard	0.065	0.220	0.546	0.595
29	bird	woodland	0.013	0.220	0.335	0.505
33	hill	woodland	0.145	0.324	0.590	0.810
37	magician	oracle	0.130	0.280	0.438	0.580
41	oracle	sage	0.283	0.324	0.428	0.575
47	furnace	stove	0.348	0.198	0.721	0.715
48	magician	wizard	0.355	1.000	0.641	0.615
49	hill	mound	0.293	1.000	0.739	0.540
50	cord	string	0.470	0.800	0.685	0.675
51	glass	tumbler	0.138	0.800	0.649	0.725
52	grin	smile	0.485	1.000	0.493	0.695
53	serf	slave	0.483	0.471	0.394	0.830
54	journey	voyage	0.360	0.800	0.517	0.610
55	autograph	signature	0.405	0.800	0.550	0.700
56	coast	shore	0.588	0.800	0.759	0.780
57	forest	woodland	0.628	1.000	0.700	0.750
58	implement	tool	0.590	0.800	0.753	0.830
59	cock	rooster	0.863	1.000	1.000	0.985
60	boy	lad	0.580	0.800	0.663	0.830
61	cushion	pillow	0.523	0.800	0.662	0.630
62	cemetery	graveyard	0.773	1.000	0.729	0.740
63	automobile	car	0.558	1.000	0.639	0.870
64	midday	noon	0.955	1.000	0.998	1.000
65	gem	jewel	0.653	1.000	0.831	0.860

Table B.1: Raw results for *title* metric testing using STSS-65.

C
Person field

Algorithm 3 Algorithm for the *person* similarity metric.

Input: Two term vectors A and B
Output:

sim $\leftarrow [|A|] * [|B|]$
compare $\leftarrow [0] * |A|$

for i from 0 to $|A| - 1$ **do** ▷ Generate Jaro-Winkler similarity matrix
  for j from 0 to $|B| - 1$ **do**
    sim$[i][j].a, b, v \leftarrow i, j,$ JAROW(A_i, B_j) ▷ JAROW calculates the Jaro-Winkler similarity metric for the two string supplied to it.
  end for
end for

for i from 0 to $|A| - 1$ **do** ▷ Sort similarity values
  sim$[i] \leftarrow$ **sortByV**(sim$[i]$)
end for

```
for i from 0 to |A| − 2 do
    for j from i + 1 to |A| − 1 do
        k ← compare[i]
        l ← compare[j]
        m ← sim[i][k]
        n ← sim[j][l]

        if m.b = n.b then
            if k + 1 < |A| and m.v < n.v then
                compare[i] ← compare[i] + 1
            else if l + 1 < |A| and m.v > n.v then
                compare[j] ← compare[j] + 1
            else if m.v = n.v then
                if k + 1 < |A| and sim[i][k + 1].v < sim[j][l + 1].v then
                    compare[i] ← k + 1
                else if l + 1 < |A| and sim[i][k + 1].v > sim[j][l + 1].v then
                    compare[j] ← l + 1
                else if k + 1 < |A| then
                    compare[i] ← k + 1
                else if l + 1 < |A| then
                    compare[j] ← l + 1
                end if
            end if
        end if
    end for
end for

s ← 0
for i from 0 to |A| − 1 do
    matches[i] ← sim[i][compare[i]]
    a ← matches[i].a            ▷ Save the A, B element matches that were just found
    s ← s + |A_a|               ▷ Find the combined length of all elements in A
end for
s ← 1/s

result = 0
for i from 0 to |A| − 1 do
    a ← matches[i].a
    v ← matches[i].v
    result ← result +(|A_a| ∗ s ∗ v)
end for

return result
```

D
Process field

D.1 Types

List of process types and their associated keywords

- Albumen
  - albumen
- Ambrotype
  - ambrotype
- Calotype
  - calotype
  - talbotype
- Salted paper
  - salted
  - silver, chloride
- Collotype
  - collotype
- Carbon
  - carbon
  - carbo
  - ozobrome
- Collodion
  - collodio
  - wet, plate
  - oxymel
- Cyanotype
  - cyanotype
  - blueprint
- Daguerreotype
  - daguerreotype
- Gelatin silver
  - gelatin, silver
  - bromide
  - silver, printing, out, paper
  - gelatino, chloride
  - dry, plate
- Gum print
  - gum
- Kallitype
  - kallitype
- Platinum print
  - platinum
  - palladium
  - platinotype

    - palladiotype
- Tintype
    - tintype
    - melainotype
    - ferrotype
    - tin, type
- Woddburytype
    - woodburytype
    - woodburygravure
- Lithograph
    - lithograph
- Albertype
    - albertype
- Halftone print
    - halftone
    - letterpress
- Photochrom
    - photochrom
- Bromoil print
    - bromoil
- Photogravure
    - photogravure
- Uranium print
    - uranium
    - wothleytype
- Transparency
    - lantern, slide
    - transparency
    - transparencies
    - autochrome

D.2 Groups

List of process groups and their associated keywords

- Paper positives
    - photographic, print
    - photo, print
    - photomechanical
- Paper negatives
    - paper, negative
- Glass negatives
    - glass, negative
- Direct positives

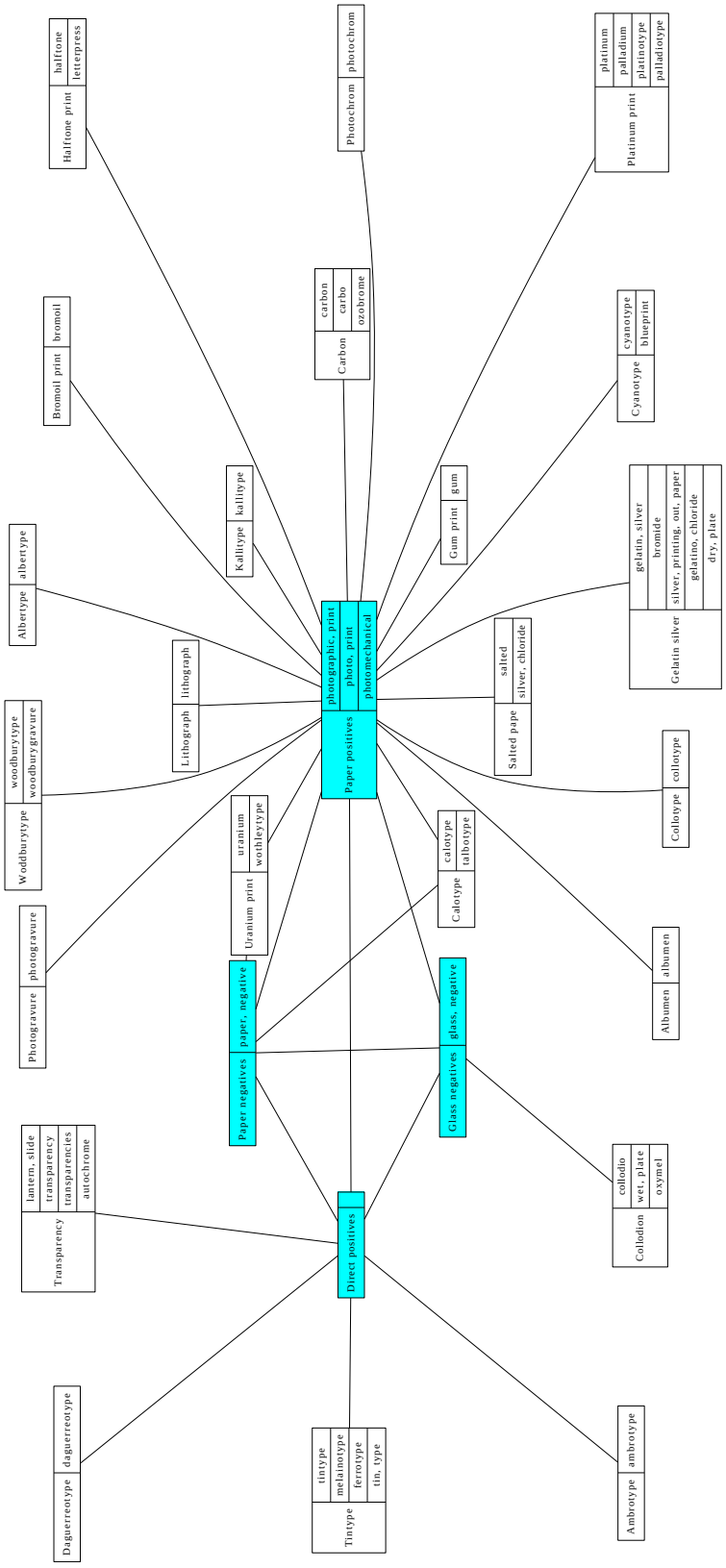


Fig. D.1: Network diagram showing the individual photographic processes, their keywords and relationships as understood by the *process* metric.

E
Overall record

Algorithm 4 Constrained minimum spanning tree algorithm.

Input: Set of records R and starting (seed) record S
Output: List of edges E describing a spanning tree for R

```
sim ← []
U ← R ⊖ S                                   ▷ Unvisited records
V ← S                                       ▷ Visited records
E ← {}                                      ▷ Record graph edges

for a from 0 to |R| − 1 do                  ▷ Generate similarity matrix as list of values
    for b from a to |R| − 1 do
        sim ← sim ∪(R_a, R_b, RecordSimilarity(R_a, R_b))
    end for
end for

sim ← sort(sim)                             ▷ Order the matrix by similarity value

while |U| > 0 do
    for i from 0 to |sim| − 1 do
        a, b, v ←sim_i

        if a ∈ V or b ∈ V then              ▷ Both nodes already visited
            sim ← sim ⊖ sim_i
        end if

        if a ∈ V and b ∉ V then             ▷ a visited but b unvisited
            U ← U ⊖ b
            V ← V ∪ b
            E ← E ∪ (a, b, v)               ▷ Set b as child of a
            break
        else if a ∉ V and b ∈ V then        ▷ a unvisited but b visited
            U ← U ⊖ a
            V ← V ∪ a
            E ← E ∪ (b, a, v)               ▷ Set a as child of b
            break
        end if
    end for
end while
return E
```

F
Testing participant responses

Uid	Manual	Test	Both	Neither	Participants
erps16243	0	0	0	1	1
erps16294	0	0	0	1	1
erps16325	0	0	0	1	1
erps16410	0	0	0	1	1
erps16470	0	0	0	1	1
erps16474	0	0	0	1	1
erps16494	0	0	0	1	1
erps16542	0	0	0	2	2
erps16545	0	0	1	0	1
erps16578	1	0	0	0	1
erps16939	0	1	0	0	1
erps16942	0	0	0	1	1
erps17093	1	2	2	2	7
erps17202	0	0	0	1	1
erps17743	0	0	0	1	1
erps18559	0	1	0	0	1
erps18912	0	1	0	0	1
erps19315	0	0	0	1	1
erps20417	0	0	0	1	1
erps20653	0	0	0	1	1
erps22202	0	0	0	1	1
erps28409	1	2	4	0	7
Totals	3 8.6%	7 20.0%	7 20.0%	18 51.4%	
Totals per Uid	1 4.3%	3 13.0%	3 13.0%	16 69.6%	

Table F.1: Occurrences of co-referent matches per search approach.

Participant	Search	Uid	Manual	Test	
1	1	erps20417	0	0	Same
	2	erps17093	0	3	Test better
	3	erps20653	0	0	Same
	4	erps28409	9	9	Same
	5	erps22202	0	0	Same
2	1	erps16545	10	10	Same
	2	erps28409	10	10	Same
	3	erps16942	0	0	Same
	4	erps17202	0	4	Test better
	5	erps17093	0	3	Test better
3	1	erps16578	9	0	Manual much better
	2	erps17093	9	8	Same
	3	erps16542	5	4	Manual better
	4	erps28409	10	7	Manual better
	5	erps16939	0	10	Test much better
4	1	erps16542	0	0	Same
	2	erps17093	6	6	Manual better
	3	erps17743	0	3	Test better
	4	erps28409	6	10	Test better
	5	erps18912	0	8	Test better
5	1	erps16294	0	0	Same
	2	erps17093	1	6	Test better
	3	erps18559	0	8	Test much better
	4	erps28409	8	9	Same
	5	erps19315	0	0	Same
6	1	erps16325	0	0	Same
	2	erps17093	2	7	Test better
	3	erps16410	0	0	Same
	4	erps28409	7	7	Same
	5	erps16470	5	1	Manual better
7	1	erps16243	3	0	Manual better
	2	erps17093	6	6	Same
	3	erps16494	0	3	Test better
	4	erps28409	6	6	Same
	5	erps16474	0	0	Same
8	1	erps17093	2	8	Test much better
	2	erps28409	1	8	Test better
	3	erps33633	5	3	Manual better

Table F.2: Participant search approach result rankings.

G
Potential co-reference matches

See also the erps17093 candidates in table 7.2 on page 137 and the erps28409 candidates in table 7.1 on page 132.

Id	erps16545	pib30240
Source	ERPS	PEiB
Title	Mont Blanc from Argenterre	Cabinet and stereoscopic photographs of Switzerland and Savoy, taken by the wet collodion process, in four frames.
Person	William England	England, William (1816-1896)
Process	[Not Listed]	collodion, stereoscopic
Date	1895	1865
Image		N/A
Found by	N/A	Manual

Table G.1: Co-reference candidates for erps16545.

[1]Whilst this record can be found using the main BkM website, it does not appear to be available via the REST interface. See section 7.1.2.1 for an explanation.

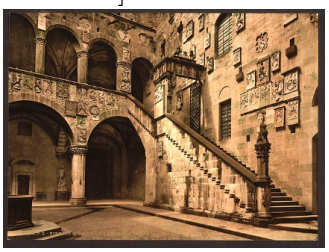
Id	erps16578	loc2001700793	N/A[1]
Source	ERPS	LoC	BkM
Title	The Courtyard of the Bargello, Florence	[Royal Museum, the court (i.e. Bargello Museum, the courtyard), Florence, Italy]	Bargello, Florence, Italy, 1895
Person	Henry Little		
Process	Bromide (Print)	1 photomechanical print : photochrom, color.	
Date	1895	[between ca. 1890 and ca. 1900].	
Image			
Found by	N/A	Manual	Manual

Table G.2: Co-reference candidates for erps16578.

Id	erps16939	erps22432
Source	ERPS	ERPS
Title	A Dutch Peasant	A Dutch Peasant
Person	James A. Sinclair	James A. Sinclair
Process	[Not Listed]	[Not Listed]
Date	1896	1903
Image		
Found by	N/A	Test approach

Table G.3: Co-reference candidates for erps16939.

Id	erps18559	erps23266
Source	ERPS	ERPS
Title	Market - Chipping Campden	The High Street, Campden
Person	W. T. Greatbatch	W. T. Greatbatch
Process	Carbon (Print)	[Not Listed]
Date	1898	1904
Image		
Found by	N/A	Test approach

Table G.4: Co-reference candidates for erps18559.

Id	erps18912	erps18911	erps17709
Source	ERPS	ERPS	ERPS
Title	South Aisle - Ely Cathedral	Stairway to Chapter House, Wells Cathedral	In the North Choir Aisle, Ely
Person	Henry W. Bennett	Henry W. Bennett	H. W. Bennett
Process	Carbon (Print)	[Not Listed]	Platinum (Print)
Date	1899	1899	1897
Image			
Found by	N/A	Test approach	Test approach

Table G.5: Co-reference candidates for erps18912.

H

Potential matches for the 'missing' ERPS photographs

This section demonstrates some potential matches found by the proposed approach for ERPS records with no image information.

Id	erps8122	loc2004676271
Source	ERPS	LoC
Title	Dandelions	Dandelions
Person	Miss Ema Spencer	Spencer, Ema
Process	[Not Listed]	1 photographic print : platinum ; 19.4 x 11.8 cm.
Date	1914	[ca. 1900]
Image		

Table H.1: Co-reference candidates for an ERPS record with no image (erps8122).

Id	erps15874	loc93510767
Source	ERPS	LoC
Title	The Lily Gatherer	The lily gatherer
Person	R. Eickemeyer, Junr.	Eickemeyer, Rudolf
Process	Platinum	1 photographic print : platinum.
Date	1894	[1892]
Image		

Table H.2: Co-reference candidates for an ERPS record with no image (erps15874).

Id	erps18923	loc2004674434
Source	ERPS	LoC
Title	Wells Cathedral; Stairs and Entrance to the Chapter House	Wells Cathedral: stairway to Chapter House
Person	F. H. Evans	Evans, Frederick H.
Process	Platinum	1 photographic print : platinum ; 9 1/4 x 7 1/2 in. (23.5 x 19 cm.)
Date	1900	1902
Image		

Table H.3: Co-reference candidates for an ERPS record with no image (erps18923).

Id	erps19389	loc2004675076
Source	ERPS	LoC
Title	The Song of the Meadow Lark	The song of the meadowlark
Person	Miss Mathilde Weil	Weil, Mathilde
Process	[Not Listed]	1 photographic print : platinum ; 18.9 x 15.4 cm. mounted on dark gray paper folder over mat, 35.9 x 26.8 cm., with cream and sage intermediate mounts.
Date	1900	[ca. 1900]
Image		

Table H.4: Co-reference candidates for an ERPS record with no image (erps19389).

Id	erps19533	loc2004676257
Source	ERPS	LoC
Title	Lady with Muff	Lady with muff
Person	Miss Mathilde Weil	Weil, Mathilde
Process	[Not Listed]	1 photographic print : platinum ; 25 x 19 cm. mounted on cream paper folded over mat, 44 x 31 cm.
Date	1900	[ca. 1900]
Image		

Table H.5: Co-reference candidates for an ERPS record with no image (erps19533).

Id	erps25184	loc2001704070
Source	ERPS	LoC
Title	The Hon. Elihu Root	Elihu Root, 1845-1937
Person	Miss Frances B. Johnston	Johnston, Frances Benjamin
Process	[Not Listed]	1 photographic print.
Date	1906	[between ca. 1890 and ca. 1910]
Image		

Table H.6: Co-reference candidates for an ERPS record with no image (erps25184).

Id	erps26130	loc93505799
Source	ERPS	LoC
Title	Feast of the Immaculate Conception	Feast of the Immaculate Conception
Person	Gertrude E. Man	Man, Gertrude E.
Process	[Not Listed]	1 photographic print.
Date	1907	c1907.
Image		

Table H.7: Co-reference candidates for an ERPS record with no image (erps26130).

Id	erps32607	loc2004676270
Source	ERPS	LoC
Title	A Mute Appeal	A mute appeal
Person	Miss Ema Spencer	Spencer, Ema
Process	[Not Listed]	1 photographic print : platinum ; 19.1 x 12.4 cm. mounted on dark gray mat, 19.4 x 12.8 cm.
Date	1914	[ca. 1900]
Image		

Table H.8: Co-reference candidates for an ERPS record with no image (erps32607).

Id	erps32622	loc97505080
Source	ERPS	LoC
Title	The Sunshine in the House	[Sunshine in the house]
Person	Mrs. Gertrude Kasebier	Kasebier, Gertrude
Process	[Not Listed]	1 photographic print : platinum ; 8 x 7 1/8 in., image 20.5 x 18 cm.
Date	1914	[1913]
Image		

Table H.9: Co-reference candidates for an ERPS record with no image (erps32622).

Id	erps33446	loc2002706463
Source	ERPS	LoC
Title	Diagonals, Brooklyn Bridge	Diagonals
Person	Arthur D. Chapman	Chapman, Arthur D.
Process	Platinotype	1 photographic print : platinum ; 8 1/16 x 6 1/16 in.
Date	1915	1913
Image		

Table H.10: Co-reference candidates for an ERPS record with no image (erps33446).

I
Questionnaires

I.1 Search technique questionnaire

Photographic collection search survey

This survey is intended to explore the use and attitudes of researchers towards online museum collections. There are a total of 11 questions. Specifically we are looking for input from those individuals involved in photographic history research although input from any individual that uses image collections for any reason is very welcome.

By completing this survey you agree for your answers to be used as part of a PhD thesis and potentially in journal and conference publications/presentations related to said thesis. The survey results will be anonymised before being used as part of any publication/presentation.

If after completing this survey you wish to have your responses removed for any reason then please contact me at david.croft@email.dmu.ac.uk.

- Person names are stored only so that responses can be removed upon request.
- If you do not supply a name, then I will be unable to remove your responses upon request.
- Name information will not be used for any other purpose.
- The non-anonymised responses will be kept in a password protected file until the thesis is complete.
- Once the thesis is completed (expected to be the later half of 2013), the non-anonymised results will be deleted along with any backups.
- Non-anonymised responses will be kept absolutely no later than 2013, the anonymised responses will be retained.

- Once the non-anonymised responses have been deleted it will no longer be possible to remove responses upon request.

Your name? _____

What is your role/interest with regards to photographic collections?* _____

Collection access

Do you use any digital collections?*

By digitised collection I mean any collection which can be search or viewed on a computer. This can include but is not limited to online collections.

◯ Yes

◯ No

How many of these collections allow you to search for specific records?*

For example, can you search by keywords? person name? date?

◯ All

◯ Some

◯ None

Searching collections

How many digital collections do you access on a regular basis?*

	1	2-3	3-5	5-10	10+
Collections	◯	◯	◯	◯	◯

When searching for specific records, what feature(s) do you consider?*

- ☐ Title/Description
- ☐ Photographer
- ☐ Photographic process (i.e. collodion negative, calotype positive etc)
- ☐ Date
- ☐ Location (i.e. the region/town/city in the photograph)
- ☐ Image size (i.e. print/negative size)
- ☐ Other: _____

When searching for specific records, what feature do you consider most important?*

- ◯ Title/Description
- ◯ Photographer
- ◯ Photographic process
- ◯ Date
- ◯ Location
- ◯ Image size (i.e. print/negative size)
- ◯ Other: _____

If you use the photographic process, are you interested in the process responsible for the negative or the positive print?

For example

- ◯ Negative process
- ◯ Positive process
- ◯ Both
- ◯ Neither, my interest is only direct positives (i.e. daguerreotypes)
- ◯ N/A
- ◯ Other: _____

Search technique

Do you ever search for a single, specific record?*

◯ Yes

◯ No

◯ Other: _____

When searching do you prefer to...*

◯ Start with a narrow focus and expand until I have the records I am looking for

◯ Start with a wide focus and narrow down until I have the records I am looking for

◯ Do both, depending on the situation

◯ Other: _____

Search results

When examining the records returned by a search...*

	1-10	10-50	50-100	100-200	200+
How many records do you typically look through?	◯	◯	◯	◯	◯
How many records would you prefer to look though?	◯	◯	◯	◯	◯
How many records are you willing to look through?	◯	◯	◯	◯	◯

Opinions

Given the digital collections that you use...

	Not	.	Ok	,	Very
When examining records returned from a search how well do you understand why those specific records were returned in response to your search?	◯	◯	◯	◯	◯
How confident are you that the records returned will be relevant to your search?	◯	◯	◯	◯	◯
How confident are you that all the potentially relevant records are returned when you search?	◯	◯	◯	◯	◯
How satisfied are you with the search systems you use?	◯	◯	◯	◯	◯

I.2 User testing questionnaire

User testing

By searching for these images and filling in the questionnaire, you agree for your answers to be used as part of a PhD thesis and potentially in journal and conference publications/presentations related to said thesis. The survey results will be kept anonymous. Your participation in this research is voluntary and you many choose not to participate. If after completing the questionnaire you wish to have your responses removed for any reason then please contact me at david.croft@email.dmu.ac.uk. Responses will be password protected and do not store any personally identifiable information. Once the thesis is completed (expected to be the latter half of 2013 it will not be possible to remove questionnaire responses from published copies of the research.

Consent*

☐ I have read the above text and consent to take part in this questionnaire.

Which collection portals did you use when searching?*

☐ The Brooklyn Museum (BKM)

- ☐ Digital NZ (DNZ)
- ☐ Exhibitions of the Royal Photographic Society (ERPS)
- ☐ The Library of Congress (LOC)
- ☐ Photographic Exhibitions in Britain (PEIB)
- ☐ The Victoria & Albert Museum (VA)

Which (if any) collection portals have you used before?

- ☐ The Brooklyn Museum (BKM)
- ☐ Digital NZ (DNZ)
- ☐ Exhibitions of the Royal Photographic Society (ERPS)
- ☐ The Library of Congress (LOC)
- ☐ Photographic Exhibitions in Britain (PEIB)
- ☐ The Victoria & Albert Museum (VA)

How long did you spend searching in total?*

_____

Search 1

Which record did you search for?*

The record ID, e.g. erps17093

_____

In your opinion, how relevant were the results found...*

	No relevance	1	2	3	4	5	6	7	8	9	Found a perfect match
...by manually searching?	○	○	○	○	○	○	○	○	○	○	○
...by the test approach?	○	○	○	○	○	○	○	○	○	○	○

How you you feel that the results from manually searching and the test approach compare?*

Manual search was much better	Manual search was better	Results were the same	Test approach was better	Test approach was much better
◯	◯	◯	◯	◯

Any additional comments?

_____

Search 2

Which record did you search for?*

The record ID, e.g. erps17093

_____

In your opinion, how relevant were the results found...*

	No relevance	1	2	3	4	5	6	7	8	9	Found a perfect match
...by manually searching?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
...by the test approach?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

How you you feel that the results from manually searching and the test approach compare?*

Manual search was much better	Manual search was better	Results were the same	Test approach was better	Test approach was much better
◯	◯	◯	◯	◯

Any additional comments?

_____

Search 3

Which record did you search for?*

The record ID, e.g. erps17093

_____

In your opinion, how relevant were the results found...*

	No relevance	1	2	3	4	5	6	7	8	9	Found a perfect match
...by manually searching?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
...by the test approach?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

How you you feel that the results from manually searching and the test approach compare?*

Manual search was much better	Manual search was better	Results were the same	Test approach was better	Test approach was much better
◯	◯	◯	◯	◯

Any additional comments?

_____

Search 4

Which record did you search for?*

The record ID, e.g. erps17093

_____

In your opinion, how relevant were the results found...*

	No relevance	1	2	3	4	5	6	7	8	9	Found a perfect match
...by manually searching?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	
...by the test approach?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	

How you you feel that the results from manually searching and the test approach compare?*

Manual search was much better	Manual search was better	Results were the same	Test approach was better	Test approach was much better
◯	◯	◯	◯	◯

Any additional comments?

_____

Search 5

Which record did you search for?*

The record ID, e.g. erps17093

_____

In your opinion, how relevant were the results found...*

	No relevance	1	2	3	4	5	6	7	8	9	Found a perfect match
...by manually searching?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	
...by the test approach?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	

How you you feel that the results from manually searching and the test approach compare?*

Manual search was much better	Manual search was better	Results were the same	Test approach was better	Test approach was much better
◯	◯	◯	◯	◯

Any additional comments?

_____

Overall

If the test approach was made available to you, would you use it for searching in the future?*

	Definitely no	No	Maybe	Yes	Definitely yes
Use	◯	◯	◯	◯	◯

Any additional comments?

_____

J
Journal/conference papers

1. Improving record matching in imprecise and uncertain datasets[45]. Journal article published in Literary and Linguistic Computing.

2. Improving record matching across disparate historical resources[46]. Submitted to and presented at the 2012 Digital Humanities Congress (DHC), University of Sheffield. It was not possible to include this paper in the thesis.

3. A hybrid approach to co-reference identification within museum collections[47]. Submitted to and presented at the 2013 IEEE Symposium Series on Computational Intelligence (SSCI), Singapore.

4. A Fast and Efficient Semantic Short Text Similarity Metric[48] Submitted to and accepted by the 2013 UK Workshop on Computational Intelligence, University of Surrey.

Improving record matching in imprecise and uncertain datasets

David Croft
De Montfort University, UK

Abstract

Museum collections represent a highly challenging search space. This article proposes a novel approach for co-referent record identification which is suitable for use across multiple separate collections. The proposed approach is intended to be suitable for use despite highly imprecise/uncertain attribute values in the records. It is hoped that this can be achieved through a combination of aspects from the fields of probabillistic record linkage, document classification, and fuzzy clustering.

Correspondence:
David Croft, 1.14 Portland Building, De Montfort University, The Gateway, Leicester, Leicestershire LE1 9BH, UK.
E-mail:
david.croft@email.dmu.ac.uk

1 Introduction

De Montfort University hosts a research database containing records of the Royal Photographic Society (RPS). This web accessible database contains the digitized contents of the exhibition catalogues produced by the RPS between 1870 and 1915 (University, 2008). It includes searchable records for the exhibited images and additional information regarding the exhibitions, competitions, judges, and awards. As a contemporary account of photography during this period, the amount of associated information makes the Exhibitions of the Royal Photographic Society (ERPS) catalogues unique. Regardless of the value of the ERPS catalogues, conspicuously absent are copies of the images being described by the records. Out of 34,197 exhibit records only 1,040 have associated images. While the ERPS catalogues already have a historical value for the photo-history community, the value could be further enhanced if the 'missing' exhibit images could be located. By identifying relationships between the entries in the ERPS database and images in collections held by other (external) institutions, the hope is that it will be possible to populate the 'missing' images.

However, the value of this research extends beyond specific information held by ERPS. The collections (both ERPS and external) within the bounds of this project exemplify a common problem in the humanities, namely, matching datasets containing imprecise and uncertain values.

2 Record Comparison Issues

When looking at the actual records to be compared, it is clear that there is no single piece of information such as 'person' or 'date', or combination of such information that can definitely identify when two (or more) separate records refer to the same item. The problem is made worse since such attributes that are available from one record or institution may not be available from another.

In addition to missing attributes, the individual attribute values have a high degree of uncertainty and can be represented in multiple ways. For example, the date attribute; since these are historical records, exact dates are not always available. This causes a degree of uncertainty when trying to compare them, i.e. is 'approx 1900' the same as '1899'? It is highly unlikely that perfect matches between attributes will occur and the partial matches which do occur result in a degree of uncertainty regarding the accuracy of the matches.

Literary and Linguistic Computing, Vol. 27, No. 4, 2012. © The Author 2012. Published by Oxford University Press on behalf of ALLC. All rights reserved. For Permissions, please email: journals.permissions@oup.com

doi:10.1093/llc/fqs028 Advance Access published on 14 July 2012

Date fields are not even the most difficult attributes to compare. In the case of the ERPS records, the most important identifying attribute is title of the image (although 'description' comes a close second when available). In the case of famous images, the 'title' field could almost be used as a unique identifier, but in the case of the ERPS records it is just a small amount of moderately descriptive text. While, metrics do exist for identifying whether two sets of text are regarding the same subject (these are mainly from the document classification domain), these typically require large amounts of text in order to produce good results.

The end result is that while there are multiple attributes that can contribute to a match, it is not possible to have complete faith in a match between individual attributes. All that can be said is that each attribute match supplies evidence towards an overall record match.

3 Existing Co-reference Methods

This concept of multiple separate records all referring to a single object is called co-reference. The important feature being that the individual records can have different or incomplete information regarding the object while still remaining valid references to it.

While identifying examples of co-reference is a vital and common part of collection curation and research, it is normally conducted at the level individual attributes; photographers, locations, and events for example. While attempts have been made at automatic co-reference identification at the attribute level (van Erp *et al.*, 2011), the evidence for automatic co-reference identification for curation is minimal (Beaudoin, 2011).

3.1 Record linkage

The obvious question is how are co-referent records currently identified? One promising area to investigate is that of record linkage (RL) (Fellegi and Sunter, 1969) which comes in two forms: deterministic (also called rules-based) RL and probabilistic record linkage (PRL). Deterministic records linkage uses a series of hand-coded rules to identify which combinations of records attributes form a set of co-referent records. This approach is both simple and fast, but can only identify co-referent records in situations that the rules designer foresaw. Despite this shortcoming, the approach is widely used, especially in industry. PRL is quite different, instead of rules each attribute of a record has a weighting value. When an attribute matches across two records, the weighting value is added to the overall match score for those two records. As more attributes match and assuming that those attributes are sufficiently weighty, the match score will exceed a pre-set threshold and the two records will be considered to be co-referent. The weighting values are produced via an analysis of a training dataset which has been pre-processed to identify the co-referent record within. The advantage of PRL over deterministic RL is that it does not require development time for the rules and can identify co-reference under unforeseen circumstances. The disadvantages are that it requires a comprehensive and representative training set for analysis and cannot identify interdependence between attributes.

3.2 Document classification

The other major approach for identifying co-reference is document classification. This is the process of grouping similar documents together according to their textual contents and can form part of search systems or simply assist in organizing documents. Regardless of intended use, there are a large number of existing methods and techniques described in the literature, however the most common theme among these techniques is the use of statistical analysis of term frequeny–inverse document frequency (TF–IDF) of keywords within the documents. A TF is simply the number of times that a specific word appears within a documents and gives a simple indication of how common said term is. This is only of limited use since it does not take into account the length of the document, longer documents produce higher values. Therefore, a TF is normally combined with a IDF value that is a general importance measure for the term across all documents in the corpus (or in the case of this project, the records). The value simply takes into

account the number of documents that the term appears in compared with the total number of documents. Combined together the result is a TF–IDF value, this gives the importance of the term within that document while taking into account the rarity of the term across the entire corpus being examined. Using TF–IDF (or similar metrics), it is possible to identify those documents with the greatest similarity to each other.

3.3 Query expansion

The major problem with the use of document classification techniques within the bounds of this project is the limited text available per record. This is further exacerbated by the records being sourced from multiple institutions, each with their own internal terminologies and due to separate writing styles and vocabularies of the individuals producing the records. The end result being that when comparing records, the number of words common to both records is expected to be very low even when they are co-referent.

Multiple synonyms being used to describe the same object is a well-known issue for both search and classification systems. The commonly used solution is query expansion and there are two major approaches. Under the first approach (global reference), the keywords in a search query are identified and are used to look up synonyms in what is effectively a digital thesaurus [often called a Lexical DataBase (LDB)]. The problems with this approach are the identification of valid synonyms and the development time required for the creation of LDB. The first occurs since the meaning of a word can change dramatically according to the context in which it is being used. The second can be addressed by using pre-existing, publicly available LDB. These have less domain-specific terminology, but this can be solved by combining a generic LDB [i.e. WordNet (University, 2011)] with a smaller domain-specific one (Mandala *et al.*, 1999).

The second approach is local feedback (Attar and Fraenkel, 1977; Croft and Harper, 1979) and has the major advantage of not requiring an LDB. However, the limited amount of text available in the records makes this approach unsuitable.

PRL demonstrates it is possible to successfully classify records without a detailed understanding of the information being classified or resorting to a rule-based approach. The information can be treated as attributes to be compared. However, the use of PRL on generic textual fields (and the imprecise attribute matches that imply) is not apparent in the existing literature. Document classification shows that it is possible to use generic textual information for classification and searching of objects. Certain approaches also demonstrate that clustering is an effective technique for achieving this (Dhillon *et al.*, 2003). However, the amount of text which is typically used in these approaches greatly exceeds the amount in most Galleries Libraries, Archives, and Museums (GLAM) catalogue records.

4 Clustering

A possible solution to this problem is clustering. Clustering is the process of grouping objects into sets based on their relative similarities. The aim being to group similar items and separate dissimilar ones. An important feature of clustering is that unlike PRL it does not require a dataset to be trained against before being able to produce results.

The methods within clustering can be divided into two major areas: partitional and hierarchical. While hierarchical clustering is the more commonly used technique for document classification, given the multi-attribute nature of the data and that partitional clustering has been shown to be effective for document classification (Steinbach *et al.*, 2000), this project intends to use partitional.

Using partitional techniques, the items being clustered can be visualized as points on a line. The relative distances between the points correspond to the degree of similarity between the items. This line analogy only functions if the items being clustered have a single attribute or are being clustered based on an overall similarity measure. As more attributes are considered, it is necessary to add additional dimensions. For example, items with three attributes being considered can be visualized as points in 3D space. The distance between two items in each dimension is the similarity measure of a single attribute while the overall distance between the two

points is the overall similarity between the two items.

4.1 Fuzzy clustering

Given the uncertain attribute values and the imprecise nature of the similarity metrics which can be used, traditional boolean clustering is not expected to produce good results. Fuzzy logic is a multi-valued logic system, as opposed to boolean logic where a statement can be either 'true' or 'false' (represented as 1 and 0, respectively). Using a fuzzy logic approach, a statement can be valid to any degree. The advantage of this approach is that it is not necessary to simplify the attribute comparisons to purely match/no-match, the results can record the uncertainty that exists regarding the matches.

Extended to clustering, the difference between fuzzy and traditional approaches is in the set membership of the items being clustered. Under a non-fuzzy clustering approach, each individual item belongs to a single set. This is acceptable for simple classification tasks. However, given the uncertainty involved in this project, caused by the generally imprecise attribute values, a non-fuzzy clustering approach is considered too restrictive and too likely to exclude valid co-referent matches for issues in a single attribute comparison. Under fuzzy clustering, every item being clustered belongs to every set to some degree (although that degree might be 0).

Fuzzy clustering has been shown to produce better classifications when compared with traditional non-fuzzy clustering approach, especially when the objects being clustered have a degree of uncertainty in their values (Mendes and Sacks, 2003). The major limitation to the greater use of fuzzy clustering is the significantly higher computational requirements which would exclude its use in any form of real-time search system. This would make it unsuitable in many areas; however, for use with the ERPS data, the low throughput does not represent a problem.

5 Proposed approach

RL, PRL, or document classification seem particularly well suited for the co-reference identification problem addressed here. RL and PRL appear unsuitable due to the textual fields of the collection records and the need for exact matches, while document classification appears unsuitable due to the minimal amounts of text found in the records. Therefore, in order to address the challenge, what is proposed is a novel combination of features from both PRL and document classification within a fuzzy clustering approach. The proposed approach keeps the individual attribute similarity value separate, this means that information regarding the comparison of each record pair is not being 'lost' and simplified into a single record similarity value. The hope is that maintaining this richer similarity information will compensate for the uncertain nature of the attribute comparison results.

Each attribute similarity measure becomes the distance between two records along one axis in n-dimensional space where n is the number of attributes being compared. The overall distance (or similarity) between two records can then be considered as the distance between two objects in n-dimensional space. As an added advantage, the proposed approach does not require simplification of the individual attribute comparison values. For example, approximate string matching methods (Damerau 1964; Winkler 1990) produce metrics which model the similarity between strings. Instead of applying static thresholds to these metrics which simplifies matching into boolean states, the proposed approach can accept the original (and uncertain) similarity metric value (assuming some processing to get the value into a [0 1] range). This allows for a fuzzy clustering approach to identify co-reference based on the actual attribute similarity (or at least the similarity metrics interpretation of it) rather than the simplified yes/no view which traditional RL requires. The same applies to the document similarity measures produced by TF–IDF which can be used to compare the textual fields of the GLAM records (i.e. the 'title' or 'description' fields).

The hypothesis is that highly similar records will be placed in the same cluster and this would include any co-referent records should they actually exist. Actual identification of co-reference will require manual examination of the results; however, the clustering process should dramatically reduce the

number of records needing to be examined compared with keyword searching.

6 Example

In order to demonstrate how the proposed approach might actually work, the initial results of the processing of record 17,654 (Table 1) from the ERPS collection are included.

The keywords identified from the title and description fields are 'chrysanthemum' and 'lady' and 'photographs'. 'The' is automatically excluded since it is an article and therefore contains no useful information, while 'photographs' is excluded since it is an exceptionally common word causing it to be a poor identifier. Using a global reference approach to query expansion [with WordNet (University, 2011) as the LDB], synonyms of these words are identified and included as an expanded word set.

Table 1 Sample record from the ERPS collection

Record Id	17,654
Collection	Erps
Title	The chrysanthemum lady
Description	Photographs
Person	Miss frances b. johnston
Process	(Not listed)
Date	1897

The expanded set is used to identify minimally matching records across all the collections by simply searching for any record which contains at least one of these words. While this method does select all records that demonstrate any resemblance to the seed record, it typically produces overly broad selections which are too large to be processed in a reasonable amount of time. This approach is just a simple keyword search system of the sort currently in use by some collections (e.g. ERPS) and will be replaced at a later date. Searching using the application programming interfaces of the Library of Congresss (LoC), DigitalNZ (DNZ), Brooklyn Museum (BkM), and Victoria and Albert museum (V&A) (Brooklyn Museum, 2012; National Library of New Zealand, 2012; L. of Congress, 2012; Victoria and Albert Museum, 2011) using this method locates 34,349 minimally matching records.

With the minimally similar records identified, it is possible to calculate the similarity matrix for each of the attributes by simply comparing every record combination. In order to visualize the similarity matrices and to demonstrate the anticipated effectiveness of the proposed approach, visual assessment of cluster tendency (VAT) images (Bezdek and Hathaway, 2002; Havens and Bezdek, 2011) have been included. A very simple example VAT image and dataset are shown in Fig. 1. In the example, twenty datapoints are organized into four clusters of varying sizes. Seen in a VAT image (shown on the

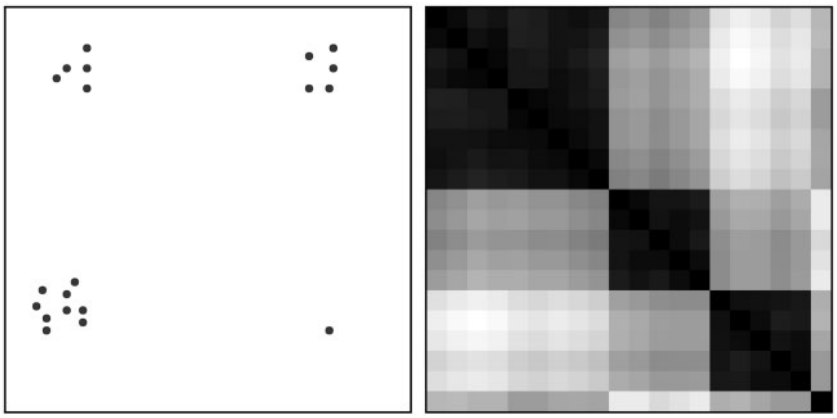


Fig. 1 Example of a VAT image with underlying dataset

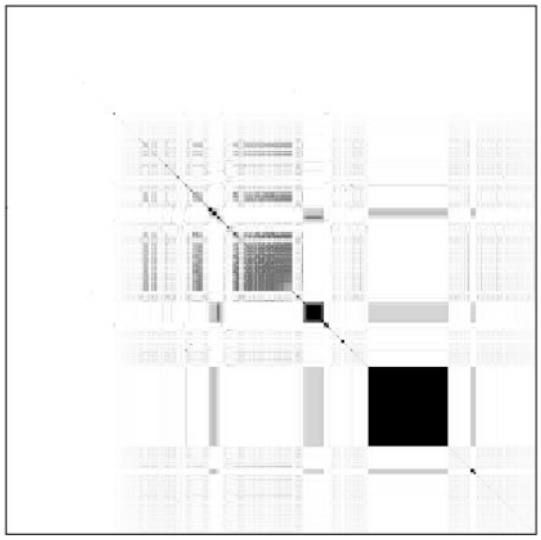


Fig. 2 VAT image of similarity matrix for description attribute of record 17,654

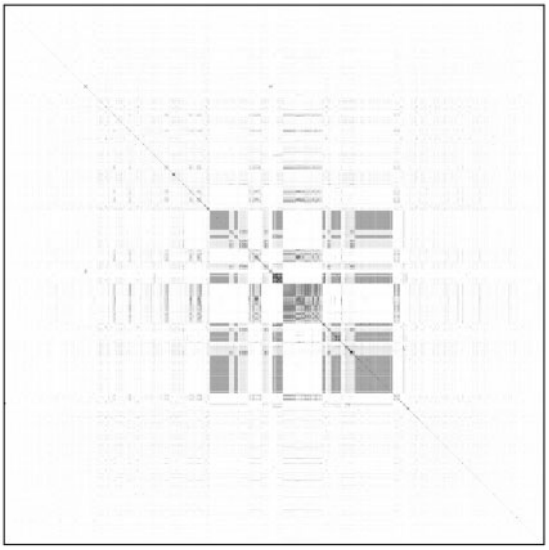


Fig. 4 VAT image of similarity matrix for combined attributes of record 17,654

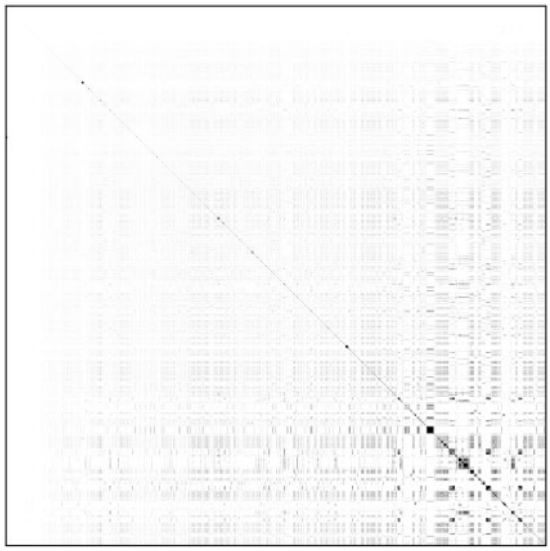


Fig. 3 VAT image of similarity matrix for title attribute of record 17,654

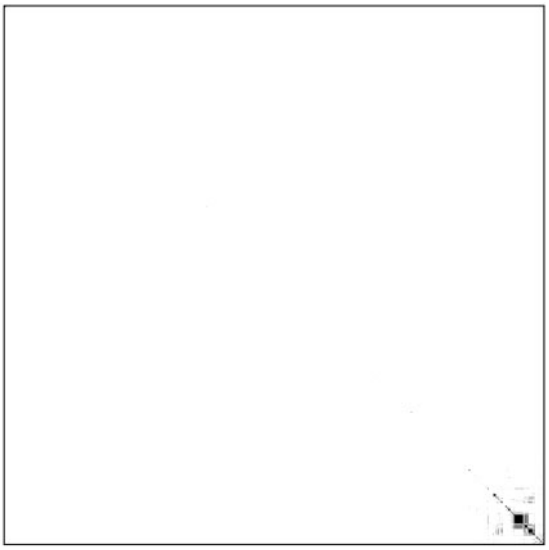


Fig. 5 VAT image of similarity matrix for date attribute of record 17,654

right), each cluster appears as a black square along the diagonal. The size of the squares corresponds to the size of the clusters. Therefore, by studying the number and size of the squares, a reasonable estimate of the number of clusters and the number of points in each cluster can be found.

The VAT images indicate the existence of clusterable structures in the similarity matrices for the title

and description attributes (Figs 2 and 3) and overall record similarity matrix (Fig. 4) which was produced by combining the matrices from the individual attributes. This is a promising indication that there are in fact clusters to be found in the data. The size and distribution of said clusters would suggest that at the very least, clustering will exclude a large number of irrelevant records which keyword searching would otherwise include.

7 Conclusion

Even at this early stage, it is possible to identify apparent clusters in the VAT images. However, as the research progresses, several improvements will be necessary. Firstly, the method for identifying the minimally matching records needs to be replaced so as to produce fewer results. Secondly, the similarity measures need to be improved (see the poor 'dates' performance, Fig 5). Finally, the inclusion of additional attributes (i.e. process used and person) and weighting the relative importance of the attributes.

Perhaps the most difficult aspect of the project will be evaluating the performance of the approach. Lacking an existing pre-classified dataset and lacking the considerable time and resources to create one, it will not be possible to evaluate performance using a quantitative methodology. Since the existence of co-reference between the collections is only an assumption and the proposed approach is believed to have applications in general record searching, basing the performance measure solely on the identification of co-referent records is overly restrictive. In the end, the performance of the proposed approach can only be measured by whether the results it produces are considered valuable by members of the GLAM community. This places the evaluation of the system in a qualitative context but the exact approach is undecided at this time.

References

Attar, R. and Fraenkel, A. S. (1977). Local Feedback in full-text retrieval systems. *Journal of the ACM*, **24**: 397–417. http://doi.acm.org/10.1145/322017.322021 (accessed 5 July 2012).

Beaudoin, J. (2011). Cluster Vision: A System to Dynamically Explore Images and Text. http://hastac.org/files/beaudoin_clustervision.pdf (accessed 5 July 2012).

Bezdek, J. and Hathaway, R. (2002). *VAT: A Tool for Visual Assessment of (Cluster) Tendency. Proceedings of the 2002 International Joint Conference on Neural Networks, 2002. IJCNN '02*, vol. 3: 2225–30.

Brooklyn Museum (2012). Brooklyn Museum. http://www.brooklynmuseum.org/ (accessed 1 January 2012).

Croft, B. W. and Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, **35**: 285–95.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, **7**: 171–6.

Dhillon, I., Kogan, J., and Nicholas, C. (2003). In Berry, M. (ed.), *Feature selection and document clustering.* Springer, pp. 73–100. ch. 4.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, **64**: 1183–1210.

Havens, T. and Bezdek, J. (2011). An efficient formulation of the improved visual assessment of cluster tendency (ivat) algorithm. *IEEE Transactions on Knowledge and Data Engineering*, **24**: 813–22.

L. of Congress. (2012). Library of Congress. http://www.loc.gov/index.html (accessed 1 January 2012).

Mandala, R., Tokunaga, T., and Tanaka, H. (1999). *Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics,* ser. EACL '99. Bergen, Norway: Association for Computational Linguistics, pp. 94–101.

Mendes, M. and Sacks, L. (2003). *Evaluating Fuzzy Clustering for Relevance-based Information Access. The 12th IEEE International Conference on Fuzzy Systems, 2003, FUZZ '03*, vol. 1. Saint Louis, Missouri, pp. 648–53.

National Library of New Zealand (2012). Digitalnz. http://www.digitalnz.org/ (accessed 1 January 2012).

Steinbach, M., Karypis, G., and Kumar, V. (2000). *A Comparison of Document Clustering Techniques. KDD Workshop on Text Mining*. Boston, MA.

University, D. M. (2008). Exhibitions of the royal photographic society 1870–1915. http://erps.dmu.ac.uk/ (accessed 1 June 2011).

University, P. (2011). WordNet: A Lexical Database for English. http://wordnet.princeton.edu/ (accessed 1 June 2011).

van Erp, M., Oomen, J., Segers, R. *et al.* (2011). *Automatic Heritage Metadata Enrichment With Historic Events. Proceedings of International Conference for Culture and Heritage On-line-Museums and the Web 2011,* Archimuse. http://conference.archimuse.com/mw2011/programs/automatic_metadata_enrichment_and_linking_fo (accessed 1 January 2012).

Victoria and Albert Museum (2011). V&A Home Page. http://www.vam.ac.uk/ (accessed 1 July 2011).

Winkler, W. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods American Statistical Association*. Washington, DC: American Statistical Association, pp. 354–9.

A hybrid approach to co-reference identification within museum collections

David Croft
School of Media and Communication,
De Montfort University,Leicester
Email: dscroft@gmail.com

Simon Coupland
Centre for Computational
Intelligence (CCI),
De Montfort University, Leicester
Email: simonc@dmu.ac.uk

Stephen Brown
Knowledge Media Design,
De Montfort University, Leicester
Email: sbrown@dmu.ac.uk

***Abstract*—Locating specific resources within museum collections represents a major challenge for users. Even when catalogues exist in a searchable digital format (which is not certain), formatting differences and the nature of the information to be encoded mean that there is a very large degree of variation in records not just from different catalogues but within individual catalogues. The nature of the data being searched means that traditional search techniques are badly suited to the challenges of identifying similar records in collections. In this paper we discuss a fuzzy rule based approach for identifying similarities between records and to identify co-referent records across multiple heritage collections. We also describe the application of this approach to real world collections and records which demonstrates some promising early results.**

I. Introduction

In a previous project De Montfort University digitised the catalogues of the Royal Photographic Society (RPS) for the exhibitions held between 1870 and 1915. The digitised information is available as a freely accessible and searchable online database. As a contemporary account of photography during this important period of development for photography in Britain the amount of associated information makes the Exhibitions of the Royal Photographic Society (ERPS) catalogues unique.

Despite the value of the collections there is one significant piece of information missing, the actual photographs. Technical limitations and the established customs with regards to exhibition catalogues at the time of the exhibitions mean that the catalogues contain images for only a fraction of the exhibited photographs. Out of 34,197 exhibited photographs the catalogues contain images for only 1,040 and many of the images are only contemporary sketches of the original photographs. The remaining images are reproductions of the original photographs printed along with the sketches in the catalogues.

We can enhance the value of the ERPS collection if visual representations for the 33,157 'missing' images and if better quality versions of the 1,040 known images were found.

With this paper we attempt co-reference identification of historical photographs using only the metadata from other distinct collections. A hybrid approach is used to fuse disparate similarity metrics using a fuzzy system.

The remainder of this paper is organised as follows;

II A brief description of recent developments in the Galleries, Libraries, Archives and Museums (GLAM) community.
III Overview of existing co-reference approaches.
IV Overview of the records and collections we used.
V Details of the individual field similarity metrics with worked examples.
VI Details of the overall record similarity metric.
VII Details of our approach for ordering potential co-reference matches.
VIII Our conclusion.

II. GLAM community/collections

Digitisation projects have been under way in the Galleries, Libraries, Archives and Museums (GLAM) community for decades and as a result millions of historical photographs have now been digitised. These photographs are often freely available and can be searched online. Recently there has been a move towards making these records available in computer readable formats such as eXtensible Markup Language (XML) and JavaScript Object Notation (JSON) and making the collections searchable using REpresentational State Transfer (REST) and SPARQL Protocol and RDF Query Language (SPARQL)[1]. These formats and interfaces allow third party software easy access to the collections. Whilst these interfaces vary in ease of use and functionality from institution to institution they still represent a massive improvement on previous approaches for connecting third party software to these resources (e.g. screen scraping).

We asked if, given that large digitised collections are now easily accessible; is it possible to locate copies of

the missing ERPS images in these external collections? If it was possible to locate copies of the images then it should be possible to learn what the 'missing' images look like and to see higher quality versions of the 'known' images. The number of collections which need to be searched, the number of missing images and the amount of time required to search for each image means that a manual approach has unacceptable time and resource requirements. Therefore this paper focusses on methods of automatically or semi-automatically identifying co-reference between images in multiple photographic collections. Since the missing ERPS images are lacking any visual information, we must achieve co-reference identification solely through analysis of the records' metadata.

III. Co-reference identification

The challenge faced in this paper is one of co-reference identification. That is, identifying when two distinct records are referring to the same item/person/place etc. even when the records contain different information some of the fields. This area has been subject to a significant amount of research since the problem appears in such a wide range of application areas[2, 3]. However, the nature of the records to be analysed in the case of GLAM records presents a number of challenges. Methods for identifying co-reference in collections with these issues (or at least with some combination of these issues) are not apparent in the literature.

- Format: Whilst the REST and SPARQL interfaces of the various collections do return the information wrapped in well established markup languages (i.e. XML and JSON). The actual field contents are typically just human readable strings in nonstandard formats. This lack of a standard format applies not just to records from different collections but often to different records within the same collection.
- Accuracy: The accuracy of the information for the records can also be suspect.
- Precision: Even when the information is correct its usefulness can be limited. Date information is a prime example of this, often unable to specify an image's origin to anything more precise than a specific century and on some occasions unable to achieve even that.
- Length: Much of the research conducted into co-reference identification on textual information has focussed on document classification. Whilst it is true that document classification faces similar challenges in identifying similarities in natural language texts, the techniques derived require a significantly greater amount of text for analysis than is available from GLAM records.

In the following subsections we describe three established approaches to co-reference identification, an expert knowledge approach, a supervised learning approach and an unsupervised learning approach in that order. We also describe the advantages and disadvantages of each approach with regards to this case study.

A. Rule based

This widely used approach uses a series of rules which have been developed for a specific knowledge base[4]. These rules will have been designed to identify co-referent records according to previously experienced and identified patterns. The major advantage of this approach is that it is relatively simple to implement with a high throughput (depending on the number and complexity of the rules), however this approach only works for the situations that the designers foresaw and encoded in the rules. A smaller but still a significant problem is that this approach is not well suited to uncertain matches between individual fields.

B. Probabilistic Record Linkage (PRL)

Identifies co-reference by assigning weight values to the individual fields in a record[5]. The overall record match can then be calculated by summing the weights for all the fields that match between any pair of records. If the overall value exceeds a preset threshold then the records are considered co-referent. However, this approach requires a training data set in order to generate suitable weights for the fields. The lack of an existing training data set and the time requirements for creating one from scratch mean that field weights would have to be generated manually.

C. Clustering

Most often seen employed in document classification tasks[6], clustering is the process of grouping objects into sets based upon their relative similarities. An important feature of clustering is that unlike PRL it does not require a data set to be trained against before being able to produce results but neither does it require rules identified and programmed.

IV. Search space

In order to try and identify co-referent GLAM records we collected 1.7 million records from the Brooklyn Museum (BkM)[1], DigitalNZ (DNZ)[2], Library of Congresss (LoC)[3] and the Victoria and Albert (V&A)[4]. The records were collected using the institutions' REST interfaces through a combination of keyword searching and a list of keywords extracted from the ERPS collection records and then query

[1]http://www.brooklynmuseum.org
[2]http://www.digitalnz.org
[3]http://www.loc.gov/index.html
[4]http://www.vam.ac.uk

expanded. The records were stored along with the records from the ERPS[5] and Photographic Exhibitions in Britain (PEB)[6] collections in a local database so as to avoid continuously querying the institutions' REST interfaces during this project. These collections were selected so as to provide a representative sample of a range of GLAM collections in the areas of collections size, information quality and information quantity per record.

Examining every record in the full collections of multiple institutions for each record searched (the seed record) is both unnecessary and impractical. Instead a keyword search was performed in order to identify those records in the collection which possesses any resemblance to each seed record. In a real system this search would have been conducted against the REST and/or SPARQL interfaces of the GLAM institutions but for this project was simply run against the 1.7 million records held in the local database.

Given our desire to search multiple collections, we only used keyword searching and record category filters (to select photographic records only). The capabilities of the search interfaces offered by GLAM institutions vary greatly. SPARQL endpoints offer capabilities comparable to a direct Structured Query Language (SQL) connection to the collection database whilst REST will typically offer keyword searching with the possibility of additional filters based on process, dates, location etc. which we did not make use of.

In order to select a manageable selection of records from the full collections we generate a complete list of all words used in both the *title* and *description* fields of the seed record. This list is then expanded using a combination of synonym expansion and generation of inflected forms (e.g. if the record features the word "flower" then also search for "flowers", "flowering", "flowered" etc.). Multiple keyword searches performed, one for each word in the expanded list and each search containing just one word. The vast majority of search results do not resemble the seed record to any significant degree but at this stage the intention is to achieve the highest search recall rate as possible whilst still keeping the number of results low enough that pair-wise comparison of the records remains possible. Once all the searches have been performed, duplicate records in the results are removed using the records' Unique Reference Identifiers (URIs), the aim is to produce a subset of minimally similar records which contains every possible record which could have been found using a manual keyword search of the full collections.

[5]http://erps.dmu.ac.uk
[6]http://peib.dmu.ac.uk

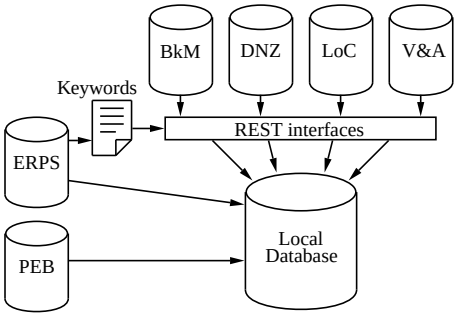


Fig. 1. Record collection to the local database.

V. Similarity metrics

Our approach is to identify the similarity between the individual fields of the records being compared before combining the field similarity metrics to produce an overall record similarity. Whilst some GLAM collection have a large number of fields for each individual record (the LoC and V&A are prime examples), the records of the ERPS collection have only five main fields. Some additional information is available in the ERPS database such as exhibition section, sales price and awards given but this is stored separately. Therefore in this report we focus on the main fields.

1) *title*: This field contains a short description of the photograph's contents (e.g. 'fair daffodils'). For the majority of records however some of the records have more poetic descriptions (e.g. 'sympathy'). Whilst these poetic descriptions accurately capture the emotional content of the photograph, they are of little use in identifying the appearance of the image.
2) *description*: This field displays the greatest variation in its contents both within ERPS and across GLAM records in general. It can contain anything from a generalised description of the photograph's contents to an in-depth technical discussion of the precise photographic processes used.
3) *person*: Contains the name of the photograph's exhibitor at the exhibition. This may also be the name of the photographer but not in all cases.
4) *process*: The chemical and mechanical process/processes which were used to create the image. The negative and/or positive process/es.
5) *date*: The year in which the photograph was exhibited in the case of the ERPS collection. In other collections it can contain the date the photograph was taken, developed or acquired by a collector/institution, it is rarely clear which.

These records represent the minimum level for most collections, at the present time we are unaware of any

collection which offers fewer fields per record although of course not all records will have all of these fields populated. We employ four distinct similarity metrics in order to compare four of the ERPS fields, only *description* is not included since this field contained such a large degree of variation between records that we were unable to reliably compare the contents between record pairs. The following sections describe the various similarity metrics employed.

A. Title

The briefness of the *title* text poses significant challenges for similarity comparison of the field. The ERPS collection records produce an average of only 5.4[7] useful words per record. As such using direct term comparison techniques such as Term Frequency (TF) would be unlikely to find any words in common between pairs of records even when said records were co-referent[7]. This means that some form of synonym expansion/comparison technique is an absolute requirement.

In preparation for comparison, the *title* fields for each pair of records are tokenised, lemmatised[8] to their stem form (using WordNet[8]) and filtered to remove low value terms (e.g. 'in', 'and', 'to' etc.) and transformed into term vectors (see table I). As mentioned previously, a comparison of terms actually listed in the fields would be unlikely to produce a good result due to the synonymy problem. In order to compare the fields based on a semantic understanding of the field text, each term vector is transformed into a weighted term vector. For each vector the weighted vector can then be produced by simply multiplying the term vector and a term similarity matrix. The initial term vectors, term similarity matrix and weighted term vector result for a comparison of 'the chrysanthemum lady' versus 'a woman selling flowers' (vectors A and B respectively) can be seen in tables I and II. The overall vector similarity can be calculated as the cosine of the two weighted vectors.

The origin of the term similarity matrix values for this project is WordNet. The individual words in the term vectors are matched to their WordNet sysnets. Synset is a single meaning of a word and so each word can have multiple synsets in order to describe the many slight variations in meaning and homonyms which can exist for a single word. Similarity values between groups of synsets were calculated using the path similarity method which as the number of nodes in the shortest path between the synsets across WordNet's

[7]34,197 records examined. Combining *title* and *description* fields gives a mean average of 8.1 words. Filtering out low value terms (i.e. 'in', 'and', 'of') produces a mean of 5.4 words per record.

[8]Normalising a word to its base form. For example 'swim', 'swam', 'swimming' and 'swims' all have the same base form of 'swim'

TABLE I

		chrysanthemum	flower	lady	selling	woman
Term	A	1	0	1	0	0
	B	0	1	0	1	1
Weighted	A	1.09	0.60	1.09	0.13	0.60
	B	0.66	1.20	0.67	1.18	1.20

TABLE II
EXAMPLE TERM SIMILARITY MATRIX

	chrysanthemum	flower	lady	selling	woman
chrysanthemum	1.00	0.50	0.09	0.06	0.10
flower		1.00	0.10	0.09	0.11
lady			1.00	0.07	0.5
selling				1.00	0.09
woman					1.00

IS-A hierarchy[9], since lower node counts indicate greater similarity the semantic similarity is calculated as $similarity = 1/nodes$. As a single term can map to multiple synsets, the similarity value used is simply the best possible match between the groups.

Despite the simplicity of this approach compared to approaches such as grammatical analysis it does succeed in scoring semantically similar *titles* with higher values than semantically dissimilar ones. Using the *title* fields examples used in table I gives an overall similarity score of 0.76. Whilst a more in-depth analysis of the *titles* would undoubtedly improve the similarity scores produced, since this technique does have problems with homonyms, the described technique achieves an acceptable level of semantic comprehension whilst also maintaining the processing throughput required for pair wise comparisons.

B. Person

Whilst name comparison is a common problem and has been mostly solved, GLAM records present a number of challenges not typically found elsewhere.

1) Name order: Depending on the record the individual elements can be stored in any order.
2) Short forms: Including initials.
3) Additional information: Mostly commonly details of birth and death years.

In our approach the name strings are first tokenized and punctuation removed to produce two vectors, each containing the elements from a single *person* field. In order to compare the individual elements of the names whilst still allowing for typos, the Jaro-Winkler[10] algorithm is used to populate a $m \cdot n$ similarity matrix where m and n are the previously mentioned vectors and $n <= m$. On average there are only 3.15 words

TABLE III
ORDERED JARO-WINKLER SIMILARITY MATRIX.

	benjamin		*frances*		*johnston*
b	0.71	frances	1.00	johnston	1.00
johnston	0.47	johnston	0.51	frances	0.51
frances	0.35	miss	0.00	miss	0.00
miss	0.00	b	0.00	b	0.00

TABLE IV
person FIELD SIMILARITY METRIC RESULT.

	benjamin b	frances frances	johnston johnston
Jaro-Winkler	0.71	1.00	1.00
Length	4.5	6	8
Weight	0.23	0.36	0.41
Combined	0.16	0.36	0.41
Result			0.93

per *person* field[9] and so full matrix generation is fast

The next stage is to find that best match for each element of n to the elements in m. We enforce match exclusivity and so each element of m can match a single element of n. The aim is to find the best overall match of the elements given this constraints. Performing an exhaustive search of every possible combination of elements is slow since there are m^n possible combinations of the elements. However, in most cases the best match for each m element will be a different element of n. This means that by ordering the elements of m according the Jaro-Winkler values for each element of n(see table III), the most promising combinations can be searched first, removing the need for an exhaustive search and significantly reducing the processing requirements. At this stage we have a match for each element of n to another element in m.

The Jaro-Winkler results for these matches are then weighted according to the mean of the length of the elements in each match as a proportion of the sum of the length of all elements. This stage is important as it means that matches based on initials are given lower importance than matches based on full names, a match between 'b' and 'benjamin' is clearly less valuable than a match between two instances of 'johnston'. Without this weighting, matches between the initials of two names are considered just as valuable as a match between two full surnames. The final similarity is then just the average of the weighted values.

C. Process

Of the fields we successfully produced similarity metrics for, this was the most challenging. The majority of photographs require one set of processes in order to create a photographic negative and a second set to produce the positive image from that negative. The majority of GLAM records have either the negative or the positive process listed. Additionally GLAM collections suffer from high process miss-identification rates. Whilst this problem is referred to repeatedly in the literature it does not appear that any research has been conducted to identify what this rate is, which processes are most likely to be misidentified or which processes they are misidentified as.

Given the lack of hard data on misidentification rates, this metric operates on the assumption that the processes which are most likely to be confused as one another are the processes which are most similar taxonomically. For example both the albumen and platinum processes produce a positive image on paper and are therefore more likely to be confused with one another than with the daguerreotype process since that produces positive image directly on metal plate. Therefore our similarity metric operates in a similar manner to the path finding across the IS-A hierarchy in WordNet, the difference being that hierarchy is of photographic processes and the nodes are weighted (see fig. 2).

The first stage of the metric is to compare the tokenized contents of the *process* fields against a series of identifying keywords for each process using Jaro-Winkler. For example the tintype process is associated with the keywords, "tintype", "ferotype" and "melainotype". A *process* field is associated to the best overall match across all the keyword lists only if the overall match exceeds 0.85. Jaro-Winkler is used for the keyword comparison to ensure that minor typos and spelling variations do not prevent matches. If the fields match at this process level then a similarity of 1.0 is achieved. The hierarchy is only four levels, at the top "direct positive", "paper positive", "paper negative" and "glass negative". The *process* similarity is the lowest cost path between the processes listed in each of the fields compared across the weighted edges.

This main flaw in this approach is that the contents of the hierarchy were generate manually as were the weights assigned to the edges. The hierarchy is by no means an exhaustive list of all photographic processes, the processes listed are only those actually mentioned in the 1.7 million locally stored records. Improvement of the weight values would be dependent on future research being conducted by the photo history community. As it is we believe that this is the first attempt to model historical process misidentification rates when searching photographic collections.

D. Date

Whilst the *date* field of the ERPS collection always describes a single year, the *date* fields of records taken from other collections are more challenging. Ignoring

[9]452,834 of 875,267 LoC records with a non-null *person*.

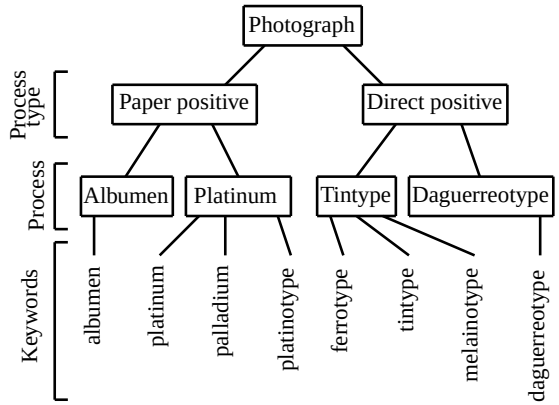


Fig. 2. Subset of the processes hierarchy.

for the moment the multitude of different formats used, many of the records describe date ranges rather than a single specific point in time. An analysis of 875,267 records collected from the LoC produced an average date range of 4.7 years per record[10]. As such it is necessary to take into account not just the distance between the dates but also the described time spans. Large date ranges that overlap perfectly may have a lower similarity than short time spans separated by a few years.

The contents of the *date* field are extracted using a rule based system which can handle most common date formats. However, a certain level of uncertainty is unavoidable for two digit years. Once the dates are extracted, the *date* metric uses three factors to produce the similarity values, the date ranges' span, start gap and end gap. These values are calculated using the year information of the dates provided. In a few cases it would be possible to produce a much narrower focus down to the individual day that the photograph was taken but only for a fraction of the records (0.004%[11]). Span is calculated as the average length of the date ranges being compared. Start and end gaps refer to the length of time between the starting/ending of one date range and the starting/ending of the other. The time span, start gap and end gap values are all scaled to a $[0\ 1]$ range, where values of ≥ 50 years have a scaled value of 0. The final date similarity is the mean of these three factors.

For example, two date ranges. A = "1888 to 1910" and B = "1874 to 1897" (see fig. 3). This produces start and end gaps of 14 and 13 years respectively, and an overall span of 36 years. The weights for these values are therefore 0.72, 0.74 and 0.28 respectively producing an overall weighting of 0.58.

A value of 0 for date ranges over 50 years may

[10]574,631 of 875,267 LoC records with comprehensible date formats. Mean average of 1,734 days. Simplified to 4.7 years.

[11]2,094 of 574,631 LoC records.

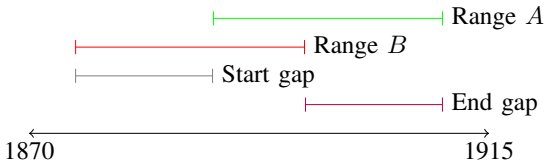


Fig. 3. Example *date* ranges and gaps.

seem low and many GLAM records have longer time spans listed (e.g. "19th century") but photography as a whole has existed for less than 200 years. Date ranges that position a photograph less precisely than a quarter of the life of the technology are not seen as useful information.

$$startgap = \min(\max(|\frac{A_{start} - B_{start}}{50}|, 0), 1) \quad (1)$$

$$endgap = \min(\max(|\frac{A_{end} - B_{end}}{50}|, 0), 1) \quad (2)$$

$$span = \frac{1}{2}((A_{end} - A_{start}) + (B_{end} - B_{start})) \quad (3)$$

$$sim = 1 - (\frac{1}{3}(startgap + endgap + span)) \quad (4)$$

VI. Overall Record Similarity

It is not possible to identify record co-reference using a single field. Matching records can only be identified via an analysis of multiple features, for this reason we employ a fuzzy rule based system in order to fuse the results from the individual field similarity metrics and produce an overall record similarity value. It became clear during our research that certain fields were significantly more important than others in determining co-reference. This had been anticipated and it was hoped that a field weighting approach similar to that seen in PRL would be sufficient. Unfortunately once this approach was implemented and testing began it became clear that this approach was not able to handle the real complexities of the overall record matches. For example, without a good *title* or *person* match the overall record match must be considered bad even if the other two fields are perfect matches. Then again if only the *title* and the *person* matches are good then the overall match should be good regardless of the other fields. The full set of fuzzy rules used is as follows;

IF *title* is good **AND** *person* is good **THEN** match is good.

IF *title* is good **AND** (*date* is good **OR** *process* is good) **THEN** match is ok.

IF *person* is good **THEN** match is ok.

IF *title* is bad **AND** *person* is bad **THEN** match is bad.

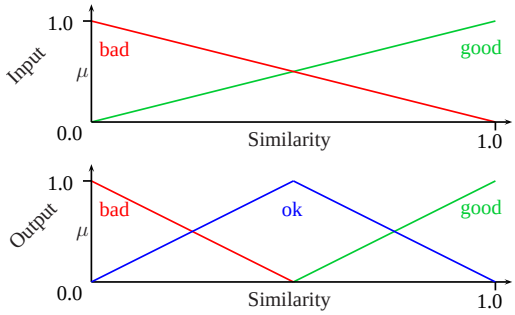


Fig. 4. Fuzzy sets used for overall record similarity.

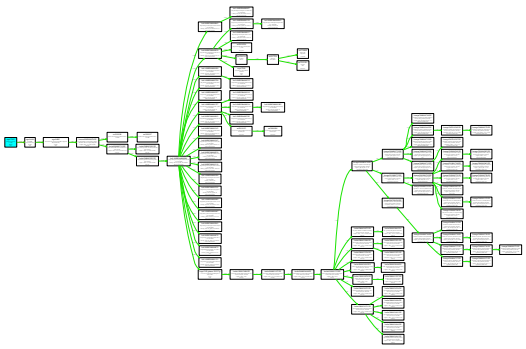


Fig. 5. Example result, top 100 results for ERPS record 17093.

The membership functions and rules of the fuzzy system are quite simple. Membership of the 'good' match input set has a linear relationship to the similarity metric values (see the green set in fig. 4). Potentially this step could be skipped and the direct similarity metric values could be used instead of the output from the fuzzification since both are in the range $[0\ 1]$, but should a nonlinear relationship between the similarity values and the fuzzified values be required then this approach can provide that. As it is, the input membership sets shown in fig. 4 are used for all four fields considered. The output sets (see fig.4) are designed to output values across the whole $[0\ 1]$ range.

Whilst the fuzzy rule based approach is much slower than other approaches, notably the sum of the similarity values as used by PRL, the throughput is sufficient. The major issue with this approach is the need to calculate pair wise similarity values among all the records. The record similarity values are non-directional and so a full similarity matrix of n^2 values can be simulated using only $\frac{1}{2}(n^2 + n)$ values but as the number of records increases this represents a major problem in both processing time and storage capacity.
dat

VII. Record Selection

With the overall record similarity matrix in place we are able to extract promising connections from the data in order to identify the most likely co-reference records. The approach we use is effectively a constrained breadth first search in which records are connected to their nearest neighbours in a hierarchical structure with the seed record as the root node (see figs. 6 and 5). Importantly this approach allows for records to be connected via intermediary links. For example once a single image by a person has been connected to the seed record the rest of that individual's work will typically be connected via that same connection, automatically organising the results by photographer.

In each iteration of the loop the record pairs are checked in order of decreasing similarity. The first record pair with only one of the records marked as visited is selected. The unvisited record is then marked as visited and the process repeats. Each time a new record is marked as visited, the visited and recently unvisited record are recorded as parent and child nodes in a hierarchical tree, which allows the connections between the matches to be analysed later. Marking the seed record as visited before processing the rest of the records means that the seed record becomes root node of the tree. The most similar (and therefore most likely to be co-referent) records are grouped at the top of the tree with the seed record. The algorithm for this process can be seen in algorithm 1.

The records have a tendency to group according to their originating collection. This issue is caused by differences in the field formats and terminology used by different institutions. The similarities in terminology of field formats between records from the same collection eventually overpower the ability of the similarity metrics to identify the underlying meaning of the records' contents. This only affects the lower regions of the tree which contain the lower quality record matches which are of less interest.

We must highlight that this approach does not perform the final co-reference identification. The limited amount of information available per record means that fully automated co-reference identification is not possible at the present time. What our approach does is to order the records according to their overall similarity. This highlights which records out of all the records returned from the initial keyword searching are most likely to be co-referent. Actually stating which records refer to exactly the same photograph requires a manual examination by a domain expert.

VIII. Conclusion

This paper demonstrates our approach for identifying co-referent records in GLAM collections. The ability to automatically select the most promising records

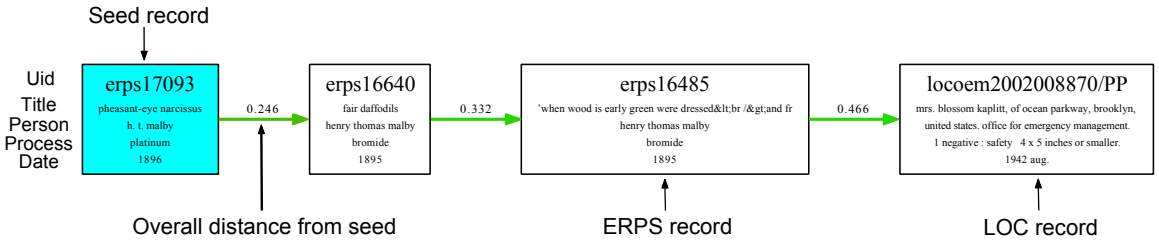


Fig. 6. Example result, top 4 results for ERPS record 17093.

Algorithm 1 Record hierarchy creation algorithm

Input: Records to be compared, records = $\{R_1, R_2, ..., R_n\}$
Onput: Hierarchical co-reference structure

```
for i = 1 to number of records do
    for j = i + 1 to number of records do
        dist ← overall similarity of records_i to records_j
        add (records_i, records_j, dist) to distances
    end for
end for

sort distances by dist

while unvisited records do
    for all distances do
        i, j, dist ← current distance

        if i is visited and j is visited then
            remove current distance from distance
        else if (i xor j is visited) and (i xor j is unvisited)
then
            if i is visited then
                set j as child of i
            else
                set i as child of j
            end if

            remove current distance from distances
            leave for loop
        end if
    end for
end while

return record hierarchy
```

based on overall record similarity is an important step towards a fully automated approach. An important feature is that our approach uses the existing record metadata in its human readable format. This means that this approach can be used to process existing collections. User input is still an necessary part of the process for sanity checking the initial search keywords and for identifying the co-referent records from the final results. Our approach reduces the number of records which need to be individually examined by the user from thousands (as produced by keyword searching) down to tens or hundreds. Reducing the number of records which need to be examined massively reduces the amount of time required to search for each of the 'missing' images in the ERPS collection. As a result our approach makes a comprehensive search for the 'missing' ERPS images a more realistic proposition as well as having wider applications for searching GLAM collections which are currently using standard keyword searching.

References

[1] A. R. Planning and R. Committee, "2012 top ten trends in academic libraries," *College & Research Libraries News*, vol. 73, no. 6, pp. 311–320, June 2012.

[2] V. Ng, "Supervised noun phrase coreference research: The first fifteen years," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1396–1411.

[3] P. Elango, "Coreference Resolution: A Survey," University of Wisconsin Madison, Tech. Rep., 2005.

[4] W. E. Winkler, "The state of record linkage and current research problems," Statistical Research Division, U.S. Census Bureau, Tech. Rep., 1999.

[5] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, 1969.

[6] A. Kyriakopoulou, "Text classification aided by clustering: a literature review," *Tools in Artificial Intelligence*, pp. 233–252, 2008.

[7] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855e, Tech. Rep., 2003.

[8] (2012, October) About wordnet. Princeton University. [Online]. Available: http://wordnet.princeton.edu/

[9] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity: Measuring the Relatedness of Concepts," in *Demonstration Papers at HLT-NAACL*, 2004, pp. 38–41.

[10] W. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage." 1990.

A Fast and Efficient Semantic Short Text Similarity Metric

David Croft*, Simon Coupland†, Jethro Shell ‡, and Stephen Brown§
* Knowledge Media Design, De Montfort University, Leicester LE1 9BH, United Kingdom
Email: david.croft@email.dmu.ac.uk
† Centre for Computational Intelligence, De Montfort University, Leicester LE1 9BH, United Kingdom
Email: simonc@dmu.ac.uk
‡ Knowledge Media Design, De Montfort University, Leicester LE1 9BH, United Kingdom
Email: jethros@dmu.ac.uk
§ Knowledge Media Design, De Montfort University, Leicester LE1 9BH, United Kingdom
Email: sbrown@dmu.ac.uk

***Abstract*—**

The semantic comparison of short sections of text is an emerging aspect of Natural Language Processing (NLP). In this paper we present a novel Short Text Semantic Similarity (STSS) method, Lightweight Semantic Similarity (LSS), to address the issues that arise with sparse text representation. The proposed approach captures the semantic information contained when comparing text to process the similarity. The methodology combines semantic term similarities with a vector similarity method used within statistical analysis. A modification of the term vectors using synset similarity values addresses issues that are encountered with sparse text. LSS is shown to be comparable to current semantic similarity approaches, LSA and STASIS, whilst having a lower computational footprint.

I. Introduction

De Montfort University hosts a research database containing records of the Royal Photographic Society (RPS). This web accessible data contains the digitised contents of the exhibition catalogues produced by the RPS. The Exhibitions of the Royal Photographic Society (ERPS) catalogues are a contemporary account of photography during the period 1870 to 1915. They hold 34,197 records but only 1,040 associated images. Whilst being of significant interest to the photo-historical community, the catalogue can be enhanced by identifying possible missing images. A wider goal of the authors work is to populate the images by comparing meta-data from external digitised catalogue sources from associated Galleries, Libraries, Archives and Museums (GLAMs) with data within ERPS. A required element of this process, is the use of Natural Language Processing (NLP), more notably semantic similarity to help match meta-data across collections [1].

There is a large body of inter-disciplinary work looking at how human language can be processed by machines in such a way that word meaning is captured in a data structure or automated process. This is generally referred to as NLP. This is a complex and dynamic goal, considered to be a discipline within Artificial Intelligence (AI) as it strives to achieve human-like performance [2].

Although the overall goal of NLP is still elusive, there have been a number of steps made towards the understanding of language. The production of parsing software [3], Part-of-speech (POS) taggers [4], [5] and Decision Support Systems (DCS) [6] have all provided inroads into the problem. However, one of the most difficult aspects of NLP is understanding semantic similarity. Humans have little difficulty in understanding the intended meaning of different words, or associating the similarity. For example, it is easy to define a level of similarity between the words *eagle* and *crane*. This maybe high if both are viewed as birds. Changing the context of *crane* to a type of machine and the similarity reduces. This is a difficult task to replicate using computation. Areas of work within similar fields, such as document classification, face similar issues when identifying similarities in natural language texts. The predominant techniques, however, require significantly greater text than is on offer within the data available to this study.

Photographic description meta-data contains sparse text. The descriptions are typically brief in length, and often grammatically incorrect, sharing many attributes with the definition of *short text* proposed by [7]. The difficulty of semantic similarity is increased when there is a reduced quantity of text. Many approaches to Short Text Semantic Similarity (STSS) [7] have been based upon existing adaptations of long-text similarity methods [8]. These methods are less applicable to our problem domain. The impact of the sentence structure and word occurrence alters with the length of text. To address these issues, we propose a novel short text Lightweight Semantic Similarity (LSS) metric. This method is compared to current approaches, Latent Semantic Analysis (LSA) and Sentence Similarity (STASIS), using a gold standard corpus.

The following sections of this paper are set out as follows. In Section II-A and II-B, the comparative methods are introduced. In Section III, an outline of LSS is given. The following section outlines the performance comparison of each method to LSS. The final section concludes the paper, discussing the findings.

II. Short Text Similarity Metrics

There are a number of approaches to measuring short text similarity. In this section we discuss two of the most popular measures, Deerwester *et al's* Latent Semantic Analysis (LSA)

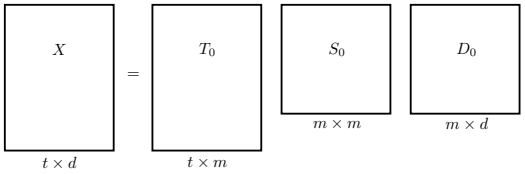


Fig. 1. Initial Matrices used in LSA.

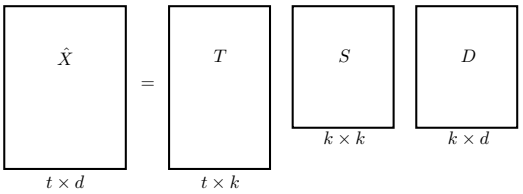


Fig. 2. Optimised Matrices used in LSA.

[9] and Li *et al's* Sentence Similarity approach (STATIS) [10]. These approaches will be compared to LSS in Section IV.

A. LSA

Deerwester *et al's* Latent Semantic Analysis (LSA) [9] is a widely used technique for comparing the similarity of short pieces of text, despite the fact that it was actually proposed for large scale data retrieval applications. LSA relates to the TF-IDF (Term Frequency - Inverse Document Frequency) approach but makes use of the singular value decomposition of TF matrices to calculate the similarity. Given d documents made up of t terms, the SVD matrices used in LSA are $X = T_0 S_0 D_0$ as depicted in Figure 1, where m is a value $\leq min(t,d)$. Redundant columns may then be removed giving a new matrix $\hat{X} = TSD \approx X$ as depicted in Figure 2, where k is number which is empirically chosen. Each row in $\hat{X}$ represents the occurrence of terms across the different pieces of text. The similarity of any two pieces of text is given by taking the dot product of two row vectors of $\hat{X}$. These can be held in a further matrix $\hat{X}\hat{X}' = TS^2T'$ where $\hat{X}\hat{X}_{i,j}$ is obtained from the cross product of row vectors $\hat{X}_i$ and $\hat{X}_j$. It is these similarity values which we are comparing the LSS method against in Section IV.

B. STASIS

1) Word semantic similarity: Similarity between individual words in STASIS is calculated as a property of relative word positions in a hierarchical knowledge base, WordNet was used in [10] but any could be used.

Terms in WordNet are represented by a set of synsets, each of which represents a differing meaning for that term. STASIS measures similarity between individual synsets using a combination of short path distance between the synsets across the WordNet's hierarchical structure and the depth of those synsets in the structure. Similarity between term pairs is calculated using equation 1.

$$s(w_1, w_2) = e^{\alpha l} \cdot f_2(h) \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (1)$$

l represents the path length between two terms and h represents the depth of the subsumer (ancestor node) in the WordNet hierarchy. α and β are tuning values, both of which should have values $\in [0,1]$. [11] used values of $a = 0.2$ and $\beta = 0.45$.

2) Sentence semantic similarity: Overall semantic similarity is calculated as the cosine of two modified term vectors. The modifications to the original term vectors made by STASIS attempt to alter identify semantic similarities between terms and to modify each term's importance in order to reduce the emphasis placed on common terms.

Semantic similarity between terms are identified as follows. For each term in the common term vector T , if the term appears in the vector (T_1, T_2) then set the value in the semantic vector to be 1. If the term does not appear in the vector (T_1, T_2), then find the term in the vector with the highest term similarity (see section II-B1), if the similarity exceeds a threshold then set the value in the semantic vector to be the term similarity. If the highest similarity does not exceed the threshold then set the value to be 0.

Term importance is identified using the information content of the terms as provided the Brown corpus [12]. The information content and the value from the previous step are combined to produce a final value for the semantic vector using the Equation 2.

$$s_i = \tilde{s} \cdot I(w_i) \cdot I(\tilde{w}_i) \quad (2)$$

The overall sentence semantic similarity is then calculated as the cosine of the two semantic vectors.

3) Word order: In contrast to LSA (amongst others), STASIS takes word order into account. This is a major distinguishing feature of STASIS when compared to other approaches which treat text as a bag of words. For example the vectors [a b c] and [c b a] are equivalent under a bag of words approach as ordering differences are ignored. STASIS however includes a computational method for measuring word order similarity between texts and so will not consider the two equivalent.

Word order similarity under STASIS is assessed as follows. The first step is to convert T_1 and T_2 into word order vectors (r_1 and r_2), This is achieved by finding the position of each term in T with that terms position in T_1 and T_2. When a term does not appear in a term vector, the position of the term with the highest similarity to the missing term is used assuming it exceeds a pre-set threshold. Otherwise 0 is used to denote position.

$$\begin{aligned} T = T_1 \cup T_2 &= [\text{a b c}] & \\ T_1 &= [\text{a b c}] & \rightarrow r_1 = [1,2,3] \\ T_2 &= [\text{c b}] & \rightarrow r_2 = [0,2,1] \end{aligned}$$

A word order similarity value can then by generated by simply calculating the normalised difference of the word order vectors (see equation 3).

$$S_r = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} \tag{3}$$

4) Overall similarity: The overall STASIS similarity for the pair of vectors being compared given by equation 4. Where $\delta \leq 1$ and controls the relative effect that the semantic similarity and word order values have on the overall text similarity value. [10] state that δ should be kept at a value > 0.5 as word order plays a lesser role in text processing[13], [10]

$$S(T_1, T_2) = \delta S_s + (1 - \delta) S_r \tag{4}$$

III. Lightweight Semantic Similarity Metric

In this Section we present a novel Lightweight Semantic Similarity (LSS) method which performs well when compared to existing approaches. This approach addresses issues when measuring textual similarity in small text sets.

The title field of a photograph is typically very short. It may also be an emotional or artistic description of the contents. The average number of title words within the description is small. The number of *useful* words is less, a mean of 5.4 words[1]. Therefore, given the briefness of the text per record, standard approaches for measuring textual similarity (such as Term Frequency (TF)) will be either unusable or will function poorly.

A secondary approach, the use of semantic meaning, additionally is problematic. The lack of sentence structure within the titles reduces the usability of the technique. The proposed methodology combines the two established approaches into a pseudo-semantic similarity with elements of the statistical techniques. The methodology combines semantic term similarities, the semantic similarity between individual terms, with a vector similarity method used within statistical analysis.

A. Text Pre-processing

The initial stage in our approach is to generate a term vector for each title field. This involves a three step process:

1) **Cleaning and tokenising each title:** The words in the title are separated and extraneous non-alphanumerical characters are removed.
2) **Removal of terms that have a high regularity:** Common terms, for example *and*, *a* and *on* commonly referred to as *stop words* within the title are extracted using the NLTK package [14]. This reduces the occurrence of high commonality words producing a high similarity measure. Many words appear frequently in searches, however, high frequency words such as *photograph*, which appears in 4% of the records collected [2], are specific to searches within the field of photographs. These are also removed.
3) **Identification of each word synset:** The synsets relating to each word are identified through the use of WordNet. WordNet is a lexical database of English words grouped into a structure of syntatic categories based on context [15]. Each word produces 0 to n synsets where n is the number of possible synsets within WordNet. Where zero synsets are identified, the raw form of the word is compared using a character based string matching. Words identified with zero synsets are maintained in the set as they can include relevant information, such as person and place names alongside technical terminology. This stage also has the effect of normalising multiple forms of the same word, for example plural, past, present and future tenses, into a single representation. This simplifies the comparisons at only a small cost to the degree of precision.

B. Similarity Metric

The pre-processing stage forms a series of term vectors that inform the similarity metric process. The term vectors represent the terms that appear in each piece of text and the number of times that each term appears in that text. Also calculated is a pair wise similarity matrix for all of the terms appearing across all pieces of text being compared. Term similarity is calculated as the maximum path similarity value (based on the shortest connecting path) between the synsets of the compared terms. This is determined by a pair wise comparison of all of the synsets corresponding to one term, with all of the synsets corresponding to the other.

The similarity metric uses a cosine similarity of the two term vectors to extract a similarity measure for the title fields being compared. Term vectors are a way to represent text and queries on the text, as vectors of identifiers. Each dimension in the vector represents a separate term. The corresponding value in the vector is non-zero, if the term appears within the text. The cosine similarities of term vectors is a common approach for identifying document similarity. Predominantly, the application of this method uses vectors contain high volume elements, hundreds or thousands. The briefness of the title fields within photographs means that is it unlikely that there will be any shared terms between pairs of titles even when they are semantically similar. Therefore the cosine similarity of the vectors will be zero.

To overcome this issue, we propose the use of a novel approach where the initial vectors are modified using the term similarity values taken from WordNet which are calculated using the method described previously. By calculating the cosine similarity on the modified, weighted term vectors, it is possible to compare according to a pseudo-semantic similarity of the terms mitigating issues caused by the shortage of text.

Cosine similarity measures the similarity of two n dimensional vectors through the use of the cosine of the angle between them. Using two elements, A and B, the similarity

[1] Combining the title and description fields for the 34,197 ERPS records gives a mean average of 8.1 words. Filtering out low values terms (for example *in* and *and*) produces a mean of 5.4 words per record.

[2] 65,491 of 1,783,280 records.

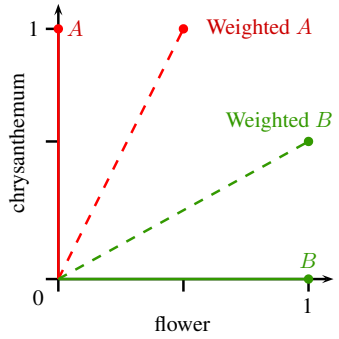


Fig. 3. Cosine of Weighted Vectors.

θ can be represented as

$$Similarity(A,B) = cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (5)$$

The results of cosine similarity produce a range of values between 0 and 1 where 0 indicates independence between the vectors and values > 0 show levels of similarity.

The limited nature of the title length can output cosine similarity values of 0, despite obvious semantic similarity. By weighting the original vectors, semantic information is incorporated. To weight the cosine similarity metric, a maximum path value is produced. The value is based on the shortest path distance needed to traverse between the two values within the WordNet tree structure. The impact of this is shown in a simple example in Fig 3.

The figure highlights two pieces of text, A = chrysanthemum, B = flower and a similarity where **sim**(chrysanthemum, flower) = 0.5. Based on the cosine similarity, the original vectors show independence as they are perpendicular to one another. By adding the weighting extracted from the WordNet distance measure, the same pieces of text produce a similarity of 0.8.

In the following section, a worked example of the proposed method will be shown.

C. Worked Example

In order to properly describe this metric a worked example of a single title pair is included. In this example the two title fields are defined as A and B, with the contents *the chrysanthemum lady* and *a woman selling flowers* respectively. Whilst the semantic similarity of A and B is obvious, there are no terms shared between the two. Therefore approaches such as TF-IDF would be ineffective. Following preprocessing of the raw fields, the original title strings produce the vectors A = [chrysanthemum, lady] and B = [flower, selling, woman]. The results of using the maximum synset similarity to generate the term similarity matrix are shown in Table I.

As the table shows, *chrysanthemum* and *flower* have a high similarity (0.50), the same applies to *lady* and *woman* (0.50), however unrelated terms such as *chrysanthemum* and *lady* have much lower values (0.09). The outcome of combining these weights with the values in the term vectors is shown in Table II.

TABLE I. EXAMPLE TERM SIMILARITY MATRIX

	chrysanthemum	flower	lady	selling	woman
chrysanthemum	1.00	0.50	0.09	0.06	0.10
flower		1.00	0.10	0.09	0.11
lady			1.00	0.07	0.50
selling				1.00	0.09
woman					1.00

TABLE II. EXAMPLE OF ORIGINAL AND CORRESPONDING WEIGHTED TERM VECTORS.

		chrysanthemum	flower	lady	selling	woman
Term vectors	A	1	0	1	0	0
	B	0	1	0	1	1
Sim matrix values for A	chry...	1.00	0.50	0.09	0.06	0.10
	lady	0.09	0.10	1.00	0.07	0.50
Sim matrix values for B	flower	0.50	1.00	0.10	0.09	0.11
	selling	0.06	0.09	0.07	1.00	0.09
	woman	0.10	0.11	0.50	0.09	1.00
Weighted vectors	A	1.09	0.60	1.09	0.13	0.60
	B	0.66	1.20	0.67	1.18	1.20

With the weighted term vectors calculated, it is possible to calculate the cosine similarity. Using the original term vectors a result of 0.00 would be achieved. However, if the similarity of the weighted vectors is calculated then a result of 0.76 is gained. We believe that this is in keeping with the semantic understanding that a human would place on the two title structures.

IV. PERFORMANCE COMPARISON

To investigate the performance of the LSS metric we ran an experiment looking at computation time and similarity compared to LSA and STASIS using a ground truth data set accepted in the literature. The performance of LSA and STASIS have already been compared by [7]. In O'Shea *et al* the results of the two approaches are compared to the averaged similarity scores from human testers using a subset of the STSS-65 dataset. In order to compare the quality of the results from the LSS metric against existing approaches, the metric was run against the same STSS-65 subset used by [7] (see table III). Figure 4 shows the LSA, STASIS and human produced similarity values plotted with the results from the LSS metric.

V. LSS METRIC TESTING

The testing data (STSS-65) consists of word pairs (see table III), whilst it more closely resembles the contents of the title fields from the ERPS collections than other data sets, it is not a perfect emulation. As such the results produced are only approximate representations of the relative performances of the tested techniques on the data we are most concerned with.

Throughput testing was conducted using Python implementations of both approaches running on an Intel Core2 Duo T5500 (1.66GHz). Alternative programming languages and/or hardware could produce faster implementations but as testing

was intended to demonstrate the comparative performance the absolute performance was unimportant.

Five sets of results were produced, the first is the time taken for LSA to produce non-directional pair-wise similarity values for increasingly large record sets. The second is the time taken for the LSS to do the same using pre-calculated word similarity values. The third is the time taken for the LSS metric if the word similarity values are not cached. Since each word pair needs only be compared once and can then be stored in perpetuity, starting with no cached word similarity values is unlikely, these results are therefore included only for completeness. Forth is the time taken by STASIS using cached word similarity values. Fifth is the STASIS time without cached values, it should be noted that the LSS metric and STASIS have different methods for calculating word similarity values, the appropriate approach was used in both cases.

A. Similarity

Adopting the approach used by [7], the LSS metric values were compared to those of the human responses using Pearson's correlation coefficient. The results of the LSS metric produced a correlation value of 0.807 compared to 0.838 for LSA and 0.816 for STASIS. This means that the LSS metric represents a performance decrease of 3.1% compared to the best performing metric LSA and 0.9% decrease compared to STASIS.

B. Computational Performance

We now consider the computation time for the metrics, again with and without cache of values where possible. Figure 5 shows the time taken for the three approaches. As can be clearly seen, the LSS metric is significantly faster than LSA when using cached results. The performance without cached values is initially worse than that of LSA but quickly improves as the number of records to compare increases. This is because the number of word similarity values which need to be calculated is directly related to the number of unique words in the records being compared. However the number of unique words per record decreases as the number of records being compared increases. Therefore computation time for LSS and STASIS compared to LSA continuously improve as more records are compared. When compared to STASIS (using pre-cached term similarity values), LSS reduces computation time by an average of 9.8%.

VI. Conclusion

In this paper we have defined the LSS short text similarity metric. This metric works by looking at the distance between synsets in WordNet and to form a term vector and then calculates the cosine similarity of this term vector. This is a simple, lightweight approach which is ideal for the problem of comparing the titles of museum artifacts, our particular problem domain.

We compared the LSS metric to two established metrics, LSA and STASIS and we found that LSS gave the best computational performance, slightly above STASIS and vastly faster than LSA and gave similarity results very close to STASIS but not as good as LSA. We believe this metric is useful as it is computationally lightweight and works well on sentence fragments of the kind found in artifact titles.

Acknowledgment

The authors would like to acknowledge the funding provided by the Arts and Humanities Research Council (AHRC) for the Fuzzy Photo Project (AH/J004367/1).

References

[1] S. C. David Croft and S. Brown, "A hybrid approach to co-reference identification within museum collections," in *CIES 2013, IEEE Symposium on Computational Intelligence for Engineering Solutions*, 2013.

[2] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[3] M.-C. De Marneffe, B. MacCartney, and C. D. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, 2006, pp. 449–454.

[4] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of international conference on new methods in language processing*, vol. 12. Manchester, UK, 1994, pp. 44–49.

[5] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 42–47.

[6] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "Methodological review: What can natural language processing do for clinical decision support?" *Journal of biomedical informatics*, vol. 42, no. 5, pp. 760–772, 2009.

[7] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A comparative study of two short text semantic similarity measures," in *Proceedings of the 2nd KES International conference on Agent and multi-agent systems: technologies and applications*. Springer-Verlag, 2008, pp. 172–181.

[8] J. Oliva, J. I. Serrano, M. D. del Castillo, and Á. Iglesias, "Symss: A syntax-based measure for short-text semantic similarity," *Data & Knowledge Engineering*, vol. 70, no. 4, pp. 390–405, 2011.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Hasrhman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391 – 407, 1990.

[10] L. Yuhua, D. Mclean, Z. Bandar, J. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 1138 – 1150, 2006.

[11] Y. Li, Z. Bandar, and D. Mclean, "An approach for measuring semantic similarity between words using multiple information sources," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 871–882, 2003.

[12] W. N. Francis and H. Kucera, "Brown corpus manual," Department of Linguistics, Brown University, Providence, Rhode Island, US, Tech. Rep., 1979. [Online]. Available: http://icame.uib.no/brown/bcm.html

[13] P. Wiemer-Hastings, "Adding syntactic information to lsa," in *PROCEEDINGS OF THE 22ND ANNUAL CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY*. Morgan Kaufmann, 2000, pp. 989–993.

[14] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, 2009.

[15] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

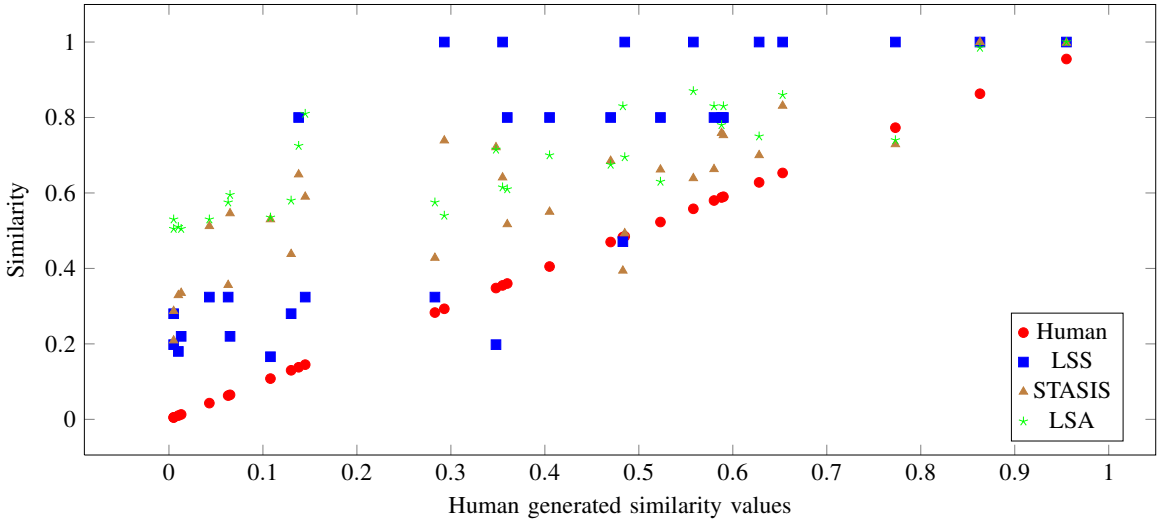


Fig. 4. Human, LSA, STASIS and LSS metric generated similarity values for STSS-65 subset.

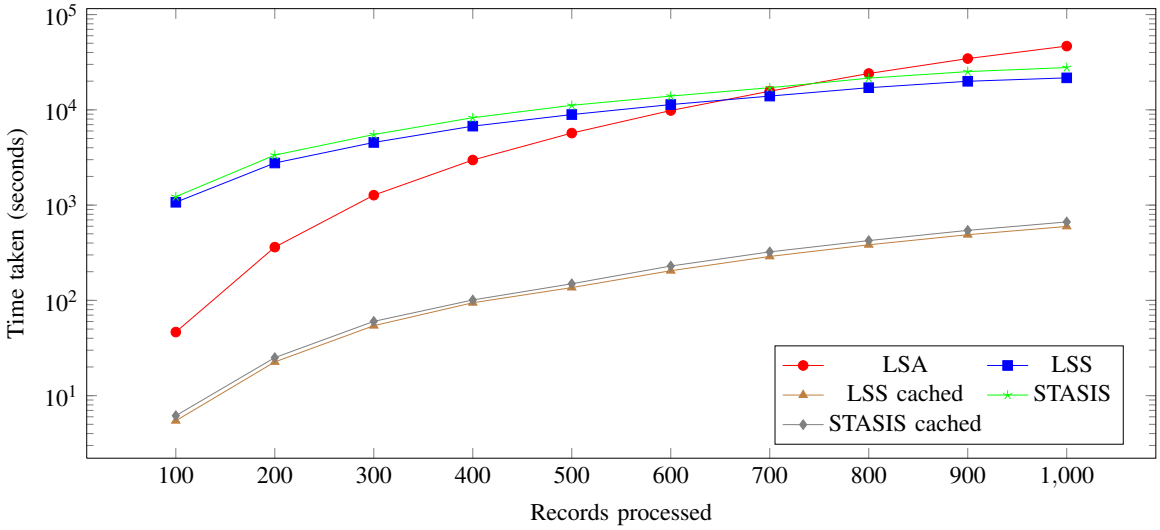


Fig. 5. Comparison of processing time requirements for LSA, LSS and STASIS.

TABLE III. RAW RESULTS FOR LSS METRIC TESTING USING STSS-65.

Sentence pair			Semantic similarity measure			
Id	A	B	Human	LSS	STASIS	LSA
1	cord	smile	0.010	0.180	0.329	0.510
5	autograph	shore	0.005	0.198	0.287	0.530
9	asylum	fruit	0.005	0.280	0.209	0.505
13	boy	rooster	0.108	0.166	0.530	0.535
17	coast	forest	0.063	0.324	0.356	0.575
21	boy	sage	0.043	0.324	0.512	0.530
25	forest	graveyard	0.065	0.220	0.546	0.595
29	bird	woodland	0.013	0.220	0.335	0.505
33	hill	woodland	0.145	0.324	0.590	0.810
37	magician	oracle	0.130	0.280	0.438	0.580
41	oracle	sage	0.283	0.324	0.428	0.575
47	furnace	stove	0.348	0.198	0.721	0.715
48	magician	wizard	0.355	1.000	0.641	0.615
49	hill	mound	0.293	1.000	0.739	0.540
50	cord	string	0.470	0.800	0.685	0.675
51	glass	tumbler	0.138	0.800	0.649	0.725
52	grin	smile	0.485	1.000	0.493	0.695
53	serf	slave	0.483	0.471	0.394	0.830
54	journey	voyage	0.360	0.800	0.517	0.610
55	autograph	signature	0.405	0.800	0.550	0.700
56	coast	shore	0.588	0.800	0.759	0.780
57	forest	woodland	0.628	1.000	0.700	0.750
58	implement	tool	0.590	0.800	0.753	0.830
59	cock	rooster	0.863	1.000	1.000	0.985
60	boy	lad	0.580	0.800	0.663	0.830
61	cushion	pillow	0.523	0.800	0.662	0.630
62	cemetery	graveyard	0.773	1.000	0.729	0.740
63	automobile	car	0.558	1.000	0.639	0.870
64	midday	noon	0.955	1.000	0.998	1.000
65	gem	jewel	0.653	1.000	0.831	0.860